Understanding the Ubiquity of Self-Deception

The Evolutionary Utility of Incorrect Information

Rauwolf, Paul

Award date:
2016

Awarding institution:
University of Bath

Understanding the Ubiquity of Self-Deception: The Evolutionary Utility of Incorrect Information

submitted by

Paul Rauwolf

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Computer Science

March 10th, 2016

Summary

When making decisions, individuals rarely possess all the facts. This can be forgiven in a world where action is time sensitive; life rarely affords the luxury of comprehending all the nuances of an environment. However, individuals do not just ignore valuable information when it is costly to acquire, individuals often ignore veridical information even when it is freely available. Instead of employing an accurate understanding of a situation, individuals frequently make decisions with the aid of ignorance and misunderstanding. This dissertation attempts to examine why. I argue against the notion that such behaviour is always caused by cognitive limitations. Instead, I demonstrate that ignoring veridical information can be advantageous in a variety of contexts. Throughout this work, I examine several settings where research has shown that individuals consistently ignore freely available information. Using a combination of formal analysis and simulations, I demonstrate that such behaviour can be advantageous. Lacking veridical knowledge can be functional in order to navigate cooperative societies (Chapter 3), unpredictable environments (Chapter 4), investment markets (Chapter 5-7), and inefficient institutions (Chapter 8). Not only does this work contribute to explaining previously confusing human behaviour, it offers insight into the potential advantages of self-deception (Chapter 2).

Acknowledgements

First and foremost, I would like to thank my father. In an unforgiving world, curiosity is easily squelched; thank you for demonstrating how to peer over the next hill. Profound appreciation goes to my supervisor, Joanna Bryson. Not only did you mould nascent ideas into works of scholarship, you patiently explained that the process is not magic. I am indebted to my colleagues Dominic Mitchell, Daniel Taylor, Swen Gaudl, and Rob Wortham. Coffee was never meant to be so enlightening. To my oldest critics, Johann Naylor and Ryan Lewis, Monday-Saturday would not have been possible without Sunday. Finally, to Rachel, for our quiet life which stands in stark contrast to your mind.

Contents

1

Introduction

> *Know thyself.*
>
> Socrates, *Phaedrus*

> *What indeed, does man know of himself! Can he even once perceive himself completely, laid out as if in an illuminated glass case? Does not nature keep much the most from him, even about his body, to spellbind and confine him in a proud, deceptive consciousness, far from the coils of the intestines, the quick current of the blood stream, and the involved tremors of the fibers? She threw away the key; and woe to the calamitous curiosity which might peer just once through a crack in the chamber of consciousness and look down and sense that man rests upon the merciless, the greedy, the insatiable, the murderous, in the indifference of his ignorance — hanging in dreams, as it were, upon the back of a tiger. In this view, whence in all the world comes the urge for truth?*
>
> Nietzche, *On Truth and Lie*

1.1 Thesis

When making decisions, individuals do not always possess an accurate understanding of the world. This is to be expected since garnering perfect information in finite time is often infeasible. As a result, temporal con-

straints are blamed for much of the fallibility in human decision-making. The assumption is that decisions would improve if information could be acquired at little cost (Shah and Oppenheimer, 2008).

Whilst this is undoubtedly the case in many contexts, my thesis is that there are a variety of, as yet undiscovered, environments where veridical knowledge hinders decision-making. In this dissertation, I analyse a variety of situations where humans are known to ignore or bias information during action selection. Rather than representing a frailty in human cognition, I demonstrate that strategies which bias or ignore veridical information can outperform individuals who use free, veridical knowledge. As such, this work extends the discussion on when biasing information is advantageous. Further, I argue that these results offer an important framework to extend the study of the evolution of cognitive biases such as self-deception and misbeliefs.

1.2 Motivation

People Make Decisions with False Information

When making decisions, individuals do not always employ a veridical understanding of the world. We generalize another's value based on a few character traits (Nisbett and Wilson, 1977). We believe we have larger control over the future (Langer, 1975) and that others agree with us (Marks and Miller, 1987) more than is objectively the case. We even hold onto inaccurate social information longer than other primates (Whiten et al., 2009).

This should not come as a surprise given the complexities of the world. No individual is going to be able to grasp all the nuances of an environment and then process the information in a Bayesian rational way (Johnson et al., 2013). Invariably, many suboptimal decisions are the result of limitations to time, information, and cognition — our rationality is bounded (Simon, 1972, 1991). Given that acquiring information takes time, and that decisions are often time-sensitive, it seems an important trade-off that we behave like many other species and "adapt well enough to 'satisfice'... not, in general, optimize" (Simon, 1956, p.129).

People Make Decisions with False Information, Even When Veridical Information is Free

There would be no further need to discuss human decision-making biases if they could be explained by temporal limitations to information acquisition. However, individuals do not just employ incorrect information when time is scarce. Individuals also tend to ignore veridical information, even when it is freely available.

For example, when making decisions with emotional consequences (i.e. affect-rich contexts), individuals tend to ignore probabilistic information, even if such information is freely available — we are probabilistically insensitive (Loewenstein et al., 2001; Pachur et al., 2014). When deciding whether to pay a cost to avoid a shock, we process a 99% and a 1% chance of a shock similarly (Rottenstreich and Hsee, 2001). When predicting future affective experiences, increasing affective intensity diminishes our sensitivity to

probability (Buechel et al., 2014). When attempting to algorithmically describe human-decision making in affect-rich contexts, ignoring probabilistic information leads to the best fit model (Suter et al., 2015).

This tendency to ignore veridical information is not just found in a few niche contexts, rather, it is ubiquitous. Despite being given an accurate historic record of a partner's social behaviour, people decide whether to cooperate with their partner based on error-prone social reputation, ignoring the veridical historic account (Sommerfeld et al., 2008). As another example, individuals systematically bias their predictions of how they will feel in a future situation (Gilbert et al., 1998). This would be excusable if someone was trying to predict how they would feel in a novel situation; however, individuals persistently exaggerate their affective predictions despite encountering the same situation time and again (Lacey et al., 2006; Scheibe et al., 2011; Wilson et al., 2003b).

Chance et al. (2011) analysed the ability of participants to predict their results on an exam. Each participant took two tests, a practice test, then a scored exam. In between the two tests, each individual predicted their results on the final exam, given their score on the practice test. The participants were divided into two groups. During the practice test, one group was given the answers to use during the exam. The control took the practice test without aid. Neither group was given any aid during the final exam. Those who had cheat sheets during the practice test predicted much higher and inaccurate results on the final exam compared to the control. The group misattributed their success on the practice test to their intelligence, rather than to the assistance of the cheat sheet.

Perhaps such a bias could be excused since the cheat sheet group had not experienced the test without the cheat sheet. It could be argued that the participants did not have enough information to deduce their own ability. Chance et al. (2015) examined this. They repeated the experiment of Chance et al. (2011) twice, in succession. First, the cheat sheet group took a test with the answers, then without, then with the answers again, and finally without any aid. The control group took four tests without any aid. Each participant predicted their result prior to each exam. After facing the second test (i.e. the first test without a cheat sheet), the experimental group was confronted with the cold reality of their unaided result. Consequently, the cheat sheet group lowered their exam predictions for the third test. However, as soon as they were given the cheat sheet again (test 3), they, once again, exaggerated their predictions of unaided exams (prediction of test 4). Despite being given sufficient opportunities to learn their actual skill, participants persisted in believing they were better than was objectively the case.

This is just one example of the human propensity to believe we are better than history reveals. Despite a life of being privy to our successes and failures, we are overly optimistic of our abilities in a variety of realms (Alicke et al., 1995; Sharot, 2011). We believe that relative to others we are better drivers (DeJoy, 1989; Dalziel and Job, 1997), have a reduced chance of divorce (Weinstein, 1980), and will live longer (Weinstein, 1980).

People Are Unaware They Ignore Veridical Information

Not only do individuals ignore freely available information, they are frequently oblivious to this behaviour. Individuals are unaware that they judge another's general nature based on a few data points (Nisbett and Wilson, 1977). Despite spending a lifetime attempting to rule objectively, judges are six times more likely to offer a prisoner parole if the judge recently ate (Danziger et al., 2011). After a food break, approximately 60% of prisoners are paroled; the percentage steadily declines until only about 10% are paroled prior to a food break. Something as simple as holding a cold versus a hot beverage decreases the likelihood that individuals will hire a prospective job candidate (Williams and Bargh, 2008).

Not only, as previously mentioned, do individuals not learn that they bias their affective forecasts, they misremember the success of their predictions. Individuals remember they are much better at predicting their emotional reactions to their team's Super Bowl loss (Meyvis et al., 2010, Study 1) and to presidential elections (Meyvis et al., 2010, Study 2) than is objectively the case. Despite iterated experience in attempting to assess our affective forecasts, we do not learn to adjust our predictions, and are "ignorant of our ignorance" (Kahneman, 2011).

1.3 Is Such Behaviour Functional?

Can ignoring free, veridical information ever confer an advantage? The null hypothesis is that such behaviours are the result of frailties in human cognition, and if individuals could learn to avoid such pitfalls, they would be better for it. This is supported by the fact that many of these biased decisions lead to harmful consequences.

Ignoring Veridical Information is Costly

Individuals who exaggerate how much discomfort they will feel in a painful situation are less likely to exercise (Ruby et al., 2011) and receive medical check-ups (Janz et al., 2007; Dillard et al., 2010). Those who exaggerate the benefits of reaching a goal waste more time pursuing impossible tasks (Greitemeyer et al., 2011). Trusting individuals are often exploited when playing economic games (Berg et al., 1995).

Ignoring Veridical Information Provides Advantages

But all is not doom and gloom. Surprisingly, there are a number of examples where biasing or ignoring veridical information appears to provide benefits. Significant evidence supports the hypothesis that positive illusions, such as self-aggrandizement, provide psychological benefits, leading to higher levels of self-esteem and mental health (Taylor and Brown, 1988). Self-affirmation is positively correlated to reduced symptoms in early-stage breast cancer patients (Creswell et al., 2007). Exaggerated beliefs in one's intelligence is positively correlated to both life-satisfaction and self-esteem (Dufner et al., 2012).

An accurate understanding of the world can be harmful. Problem gamblers more accurately predict how poorly they feel after a loss (Willner-Reid et al., 2012). Accurately predicting the enjoyment of future affective experiences is correlated to suicide attempts and escape fantasies (Marroquín et al., 2013). In general, mild to moderately depressed individuals hold a more accurate world view compared to controls (Alloy and Abramson, 1979). Further, illusion of control is inversely correlated to depressive symptoms (Alloy and Clements, 1992).

The Evolutionary Advantages of False Information

If ignoring veridical information can be propitious, could such biases have survived natural selection, providing a functional advantage? Only relatively recently has work begun to focus on whether the use of false information can confer evolutionary advantages. Whilst, as previously mentioned, it is well known that heuristics augment decision-making in time critical situations (Simon, 1956; Higginson et al., 2015; Haselton et al., 2015), it has recently been discovered that biasing information can help in navigating unpredictable (Johnson et al., 2013; Johnson and Fowler, 2011) and positively assorted environments (Fawcett et al., 2014). The evolutionary feasibility of such behaviour is further substantiated by the fact that humans share many of these decision-making biases with non-human primates (Santos and Rosati, 2015). If non-human primates possess similar biases, then perhaps such decision-making mechanisms were selected deep within our phylogenetic history.

The Evolution of Self Deception

Given that individuals are often unaware that they ignore veridical information, evolutionary theorists have begun to consider whether humans could have evolved to deceive themselves. Specifically, research has sought to understand the environmental features necessary for the evolution of self-deception and the perpetuation of false beliefs. It has been argued that self-deception can confer high levels of self-esteem (Taylor and Brown, 1988; Ramachandran, 1996), aid in deceiving others (Trivers, 1991; von Hippel and Trivers, 2011a; von Hippel, 2015), and augment decision-making in noisy environments (Ramirez and Marshall, 2015). For a full review, see Chapter 2.

Roadblock: More Data Required

Whilst such research has made preliminary progress in diagnosing when biasing or ignoring veridical information can be adaptive, the field is still nascent. Prior to attempting to generalize a theory on the necessary conditions for the evolution of ignoring veridical information, the literature must conduct a thorough search for the environments which enable such behaviour. At present, only a handful of possible explanations exist.

The same problem holds for research on the evolution of self-deception. A robust search for the selective pressures which augment self-deception is missing from the literature. How would one go about extending

the work on self-deception? Strangely, the answer to this question is rarely discussed. Rather, the conversation has focused on analysing the evolutionary feasibility of a handful of theories. As a consequence, the current literature is somewhat limited without a clear path for searching for additional environments which could promote self-deception.

1.4 Thesis: Lacking Veridical Knowledge is Ubiquitously Adaptive

My thesis is that there are a variety of, as yet undiscovered, contexts where veridical information hinders decision-making. This work begins to fill two gaps in the literature. First, it expands the known environments where ignoring or biasing veridical information is adaptive. Second, it aids in diagnosing additional environments where self-deception might be adaptive. In Chapter 2, I argue that a prerequisite for the evolution of many forms of self-deception (i.e. misbeliefs, being strategically wrong, and cognitive biases) is an environment where making a decision with false information outperforms the use of (free) veridical information. As such, by discovering additional environments which generate pressure for biasing veridical information, one is simultaneously discovering environments with potential scope for the evolution of self-deception.

Through a series of six experiments, I solve open scientific quandaries by demonstrating the advantages of employing incorrect information, even when veridical information is free. Throughout this work, I find evidence for the pervasive utility and adaptiveness of incorrectly understanding the world. Further, these findings may extend the known environments where self-deception is adaptive.

1.5 Method Overview: Adaptive Strategies

How does one evaluate whether it is advantageous to ignore information in an environment? Further, how does one demonstrate that such strategies may survive natural selection? Throughout this work, I compare the performance of strategies using differing amounts of veridical and biased information at a variety of tasks. If the dominant strategy for a given task involves biasing or ignoring veridical information, then that strategy is adaptive compared to strategies employing veridical knowledge.

Importantly, just because a particular strategy is optimal, does not of necessity demonstrate that it will evolve (Parker et al., 1990; Orzack and Forber, 2012); however, one of the basic tenets of evolution, is that "natural selection would inevitably tend to preserve those individuals which were born with constitutions best adapted to any country which they inhabited" (Darwin, 1872, p. 161). Thus, if biasing veridical information outperforms other strategies, then such behaviour is likely to be preferred by natural selection.

Throughout this work, I test for adaptive strategies via a mixture of agent-based modelling and formal analysis. I search for adaptive strategies in both frequency independent and frequency dependant environments (see below). For the remainder of this section, I discuss the methodology used throughout this work.

Definition: Veridical Information

In order to compare the utility of false versus veridical information, veridical information must be well-defined. For the purposes of this work, the correctness of information is a referent to some state of the world. For example, in Chapter 4 an agent has a real number belief (b') in the benefit of an action. However, regardless of the agent's belief, if the agent performs the action, the agent will receive some actual benefit (b). A belief which mirrors the actual benefit is considered correct (e.g. $b' = b$). Any other belief would be strictly incorrect (e.g. $b' \neq b$).

Frequency Dependent versus Independent Environments

Parker et al. (1990) delineate between searching for optimal strategies in frequency dependent and independent environments. Frequency independent environments are situations where another's strategy does not affect an individual's performance. In these environments, an optimal strategy can be uncovered by simply analysing the performance of each strategy at the given task. This can be accomplished via formal analysis (e.g. Chapter 7), or via agent-based modelling, where the performance of each strategy is simulated (e.g. Chapter 8).

In frequency dependent environments, the distribution of strategies in a population affects the performance of each individual. For example, Strategy A might be optimal in a population filled with Strategy B players, but Strategy B might outperform Strategy A when the population is equally split between A and B players. In such cases, it is important to test the performance of each strategy in a variety of population distributions. This can be accomplished via formal analysis (e.g. Chapter 3), but can also be analysed via evolutionary algorithms (e.g. Chapter 3-6).

Evolutionary Algorithms: Testing Frequency Dependent Environments

Evolutionary algorithms use simulated evolution to search through the parameter space of strategy distributions in order to find a stable strategy. To do this, an agent-based model is initiated with a certain distribution of strategies (e.g. 1% Strategy A and 99% Strategy B). The agents then perform their task, potentially competing with each other. After a period of time, each agent's score is tallied.

Next, simulated evolution is applied to create a new "generation" of agents. Since the agents who performed best are more adapted to the task, they will have a higher likelihood of seeding the next generation. As such, simulated evolution instantiates the next generation of players with a higher fraction of adaptive strategies compared to the previous generation. For example, if Strategy A outperforms Strategy B, then the initial 1% of Strategy A agents might shift to 2% in the next generation. The new generation would then simulate the task, and if Strategy A continues to outperform Strategy B, then in the subsequent generation the population may consist of 4% Strategy A agents. This would continue until an equilibrium is reached.

The flow of an example evolutionary algorithm can be written as:

1. Create a population of agents.
2. Generate variation in the population by seeding the agents with veridical and (differing amounts of) incorrect information.
3. Give the agents a task, the performance of which depends upon the varied information.
4. Simulate the agents' performance at the task.
5. Calculate the efficacy of each agent at the task.
6. Generate the next generation of agents. The next generation's distribution of strategies is biased to the strategies which performed best.
7. Repeat from item 4 until the distribution of strategies equilibrate.

Evolutionary algorithms are well suited for analysing the repercussions of frequency dependent landscapes (Alexander, 2009). Since the performance of each strategy is altered as the frequency distribution of a population shifts, evolutionary algorithms are (similar to evolution) an unsupervised search for strategic equilibria (Nowak and Sigmund, 2004). This is particularly true when modelling the effects of social interactions, where the utility of strategies are dependent upon the strategies of the other agents.

Representation

To study when false information outperforms veridical information, an agent requires the capacity to represent the information in some manner. This can be done in any of numerous ways and will vary from study to study. Generally, the particulars of the representation are not important as long as the representations of correct and incorrect information are unambiguously delineated. For instance, in Chapter 3, I analyse the reputations of players by storing the reputations in a binary matrix (i.e. an agent believes another is good [1] or bad [0]). In Chapter 4, an agent has a belief in the utility of an action. This is represented as a real number. As long as the representations of correct and incorrect information are distinguishable, then the subsequent actions caused by the representation can be acted upon by selection, and thus analysed.

1.6 Chapter Summaries

Over the course of this dissertation, I evaluate the utility of false information in solving several open questions in a variety of contexts. Here, I present a high-level overview of each chapter.

Literature Review

In Chapter 2, I review the literature on the evolutionary feasibility of self-deception. This chapter acts as motivation for continued research on the benefits of biasing veridical information. Despite significant variations in the definition of self-deception, I argue that a necessary condition for almost all definitions of the

evolution of self-deception is that the self-deceived individual must act with a lack of veridical knowledge. Given this, then the subsequent chapters expand the known environments where self-deception may be adaptive.

Value Homophily Benefits Cooperation but Motivates Employing Incorrect Social Information

When an individual discovers in isolation that a socially held view is false, should they alter their view? Empirical evidence shows that humans often do not. Individuals often judge others based on third party gossip, rather than their own experience, despite the fact that gossip is error-prone (Sommerfeld et al., 2008). Rather than judging others on their merits, even when such knowledge is free, we judge based on the opinions of third parties.

In Chapter 3, I find that conditions exist where such behaviour is adaptive. Given the importance of cooperation to humanity's evolutionary history, if truth-telling negatively impacts cooperation, lacking veridical knowledge can be adaptive. I demonstrate that signalling in-group status can outweigh honesty as the best method to ultimately spread cooperation[1].

The Evolution of the Impact Bias: Optimizing Affective Forecasts for Choice in Noisy Environments

There is robust evidence that prior to an anticipated event, people systematically fail to accurately predict their feeling after the event (Wilson and Gilbert, 2013; Wilson et al., 2000; Schkade and Kahneman, 1998; Wilson et al., 2003a; Wilson and Gilbert, 2003; Gilbert et al., 1998). They predict more intense emotions compared to their actual affective experiences, a phenomenon known as impact bias (Gilbert et al., 2002). Furthermore, individuals do not improve their predictions by learning, despite previous experience (Lacey et al., 2006; Scheibe et al., 2011; Wilson et al., 2003b).

In Chapter 4, I demonstrate that the impact bias can be functional, helping humans navigate decision-making in noisy environments. By extending the work on Error Management Theory (Johnson et al., 2013), I show that in noisy environments it can be suboptimal to learn from previous affective experiences. This explanation matches previously unexplained experimental data[2].

Trust Mediates Costly Punishment in Environments of Partial Information and Partner Choice

Note: This chapter does not directly discuss the utility of false information. However, the finding is a requisite for the results in Chapter 6 and 7.

[1]This work has been published in the Journal of Theoretical Biology. It was also accepted as an oral presentation at the European Human Behaviour and Evolution Association (EHBEA) conference in 2014 as well as the International Conference of Social Dilemmas in 2015. Further, ideas from this work have been published in the Washington Post under the title "A Tendency to Follow the Herd Rather than Whistleblow May be Part of our Evolutionary Past."

[2]This work appeared as a peer-reviewed conference paper in the Proceedings of Collective Intelligence, 2014. Further, it was accepted for oral presentation at the European Human Behaviour and Evolution Association (EHBEA) in 2015. It is currently under review.

When playing economic games, such as the Trust Game, individuals tend to make two seemingly paradoxical errors. First, when individuals possess no information about a partner's trustworthiness, they accept another's offer of unknown quality despite the fact that the subgame perfect strategy is for their partner to defect (Berg et al., 1995). Second, even when participants know a trade is profitable, they reject it unless the trade is fair — they express costly punishment against a potential benefactor (Henrich et al., 2005; Marlowe et al., 2010).

While some suggest such behaviours elucidate a human predisposition to value fairness intrinsically (Fehr and Schmidt, 1999; Fehr et al., 2002), in Chapter 5, I demonstrate that such behaviour can be revenue maximizing. Further, under the natural limitation of partial information, I show that blind trust can mediate the evolutionary feasibility of costly punishment. This result is important because it demonstrates that the evolutionary viability of trust and costly punishment may be linked[3].

Heterogeneity in Costly Punishment Strategies is Adaptive

In Chapter 5, I show that costly punishment (paying a price to punish those who refuse to treat others fairly) can be adaptive. Whilst this explains why some humans costly punish, it does not explain the variation of costly punishment strategies witnessed in experimental data (Manapat et al., 2012). If it is advantageous to costly punish, why is there such individual variation in the behaviour?

In Chapter 6, I demonstrate that heterogeneity of costly punishment strategies can be adaptive. Just as in Chapter 5, this is true assuming partner choice and partial information. Rather than learning the optimal amount of costly punishment, individuals benefit from playing a variety of strategies, never learning the best strategy. This behaviour is adaptive because third-parties react to costly punishment whether individuals punish frequently or sporadically.

Variation in Costly Punishment is Frequently Not Penalised in the Trust and Ultimatum Game

Like the previous chapter, this work considers variation in costly punishment. In recent years, individual variation in costly punishment has confused experimenters (Manapat et al., 2012). Whilst in Chapter 6 I show that such variation can be adaptive, in Chapter 7 I demonstrate that, even when variation is not adaptive, suboptimal strategies often go unpunished in economic games. If enough individuals in a population possess veridical information (i.e. knowledge of the optimal strategy), then those who lack veridical knowledge go unpunished — consequently lacking veridical information drifts neutrally.

[3]This work was accepted as an oral presentation at the International Conference of Social Dilemmas in 2015. It is currently under review.

In Inefficient Healthcare Institutions, a Mixture of Formal and Informal Reviews Leads to the Best Patient Care

In Chapter 8, I warn that inefficiencies in public institutions may generate environments where it is best to promote information obfuscation. I consider this in the context of whistle-blowing policy in the healthcare system. I demonstrate that institutional resource limitations and processing inefficiencies may make information transparency suboptimal. I analyse the utility of whistleblowing and information transparency given the efficacy of an institution. I find that given even small inefficiencies in processing inquiries, it is best for patient care if individuals either 1.) limit whistleblowing rates, or 2.) are less aware of bad practice (i.e. reduced transparency).

I conclude with a discussion on how to avoid organizational structure which (sadly) leads to pressure for information obfuscation. Utilizing simulations, I demonstrate that a mixture of formal and less formal reviews may lead to the best patient care. I conclude with a call for further research on a more holistic understanding of the interplay between organizational structure and the benefits of information transparency to patient care.

Discussion and Conclusion

In the final chapter, I discuss the implications of this work given existing research on the benefits of lacking veridical knowledge. Next, I discuss the types of environments which lead to self-deception and biases in knowledge. I conclude with a call for further and more systematic research to explain the phenomena of self-deception and the propensity to ignore veridical knowledge.

2

Literature Review

> *The mind is its own place, and in itself can make a heaven of hell, a hell of heaven.*

John Milton, *Paradise Lost*

> *A great deal of intelligence can be invested in ignorance when the need for illusion is deep.*

Saul Bellow, *Jerusalem and Back: A Personal Account*

The intention of this chapter is twofold. First, it is a literature review focusing on the evolutionary feasibility of self-deception. The phenomenon has been defined in a variety of ways, so I review several evolutionary theories which can fall under the heading of self-deception, including the evolution of misbeliefs and being strategically wrong. Through the course of this discussion, I find the literature lacks a generalizable framework for diagnosing the evolvability of self-deception. Instead of presenting a list of necessary conditions, much of the work tends to focus on analysing whether a particular variable can provide the necessary selective pressure to enable self-deception. Consequently, systematic research on the feasibility of the evolution of self-deception is lacking.

This gap in the literature leads into this chapter's second contribution — to act as motivation for researching when it is adaptive to lack veridical knowledge of the world. I argue that, regardless the definition, the evolutionary viability of self-deception depends on whether it is adaptive to lack veridical knowledge. Since this dissertation argues that there are many, previously undiscovered contexts where it is adaptive to

use false information (see Chapters 3-7), I posit this work offers insight into widening the scope of research on self-deception.

2.1 Definitions

2.1.1 Misbelief, False Belief, and Being Strategically Wrong

Prior to discussing self-deception, it is important to clarify a few related terms which are applied throughout this work. I use "misbelief" and "false belief" interchangeably. They are "a belief that is not correct in all particulars" (McKay et al., 2009). Related, is the concept of being strategically wrong. An individual is strategically wrong when it is in their best interest to possess a false belief (Kurzban, 2012).

2.1.2 Deception

The definition of self-deception is still debated (Deweese-Boyd, 2012) and hinges on other open philosophical and neurological questions regarding the concepts of identity, mind, and what it is to "know, intend, choose, and wish" (Fingarette, 1969). Even the phenomenon's constituent parts are difficult to characterize. Ignoring self-deception for a moment, the definition of *deception* is still controversial.

When has an agent been deceived? The usage of the term has diverged depending on discipline (Mitchell, 1986). For the most part, the philosophical conversation has focused on human deception, and, as a result, most agree that deception requires the deceiver to *intend* (in the folk psychology sense) to deceive another (Mahon, 2015; Fallis, 2010). For evolutionary biologists, however, intention is often not a prerequisite, suggesting that non-human species are capable of deception (Trivers, 1985), or at the very least, the functional equivalent of deception (Dawkins, 1976).

Intentional Definition

The importance of intention to human-level deception is nicely summarized by Mahon (2015). When does agent A deceive another agent B? A simple definition might only require that agent A causes agent B to believe a falsehood. More formally, because of Agent A, Agent B believes some proposition $\neg p$ when in fact p is true. For some, this definition is too broad, as it includes individuals who do not intend to impart any information. However, adding the requirement that Agent A must intend to impart $\neg p$ to agent B, does not circumvent the issue. Agent A could intend to instil $\neg p$ while believing $\neg p$ to be true. A more robust definition requires agent A to intend to cause agent B to believe what Agent A believes to be false (i.e. $\neg p$) — agent A believes p (for a more protracted analysis of this summary see Mahon, 2015).

It is important to note, however, that the need for intention in a deceptive act is highly controversial. Requiring intention traps deception in an anthropocentric box. If intention and knowledge of falsity are requisite for deception, what does that mean for other species (Bond and Robinson, 1988a)?

Functional Definition

Deception is frequently used within evolutionary biology to describe instances where species appear to have evolved to induce errors in others (Mokkonen and Lindstedt, 2015; Bond and Robinson, 1988b; Trivers, 2011). Smith (2004) defines deception as "any form of behaviour the function of which is to provide others with false information or to deprive them of true information" (Smith, 2004, p.14). Admittedly, defining the function of a behaviour is challenging (McKay et al., 2009), but natural selection helps. Without getting bogged down in debating the definition of "function", I define a process's function as one that aids in the individual's replication (Millikan, 1989). Functional deception would then be a process where inducing errors in others aids in the process's evolutionary survival.

By this definition, examples of functional deception in non-human species are seemingly endless. Floral mimicry is ubiquitous (Schiestl and Johnson, 2013), whereby the colouration of flowers lure in pollinators by imitating the pollenators' prey (Jin et al., 2014) or sexual partners (Gaskett, 2011). Prior to mating, male nursery web spiders often offer insects wrapped in silk to females. However, $1/3$ of all males offer worthless, silk wrapped plant fragments, but succeed in copulation prior to detection (Ghislandi et al., 2014). Not only do fork-tailed drongos fake alarm calls to steal food, they vary their alarm calls so that the deceived does not habituate to the false alarms (Flower et al., 2014).

2.1.3 Self-Deception: A Philosophical Soiree

As the definition of deception is still debated, such woes inevitably filter into understanding self-deception. How does one deceive oneself? It seems natural to attempt to retain the definition of deception regardless of whether the deception is intra or interpersonal. If deception requires intention (see 2.1.2: Intentional Deception), then agent A, who believes p, attempts to convince agent B to believe $\neg p$. Self-deception only requires one additional constraint — agent A = agent B (for an early review see: Kipp, 1980).

However, this gives rise to both the static and dynamic paradoxes of self-deception (Mele, 1997). The static paradox questions how the mind can both believe and not believe something. If agent A knows p and succeeds in convincing itself of $\neg p$, does it still know p? Can an agent believe both p and $\neg p$? The dynamic paradox argues that if deception must be intentional (in the folk psychology sense), then how can deception succeed if the one being deceived knows they are being deceived (Mele, 2001)? Historically, there are two paths for attempting to overcome these hurdles:

1. the intentionalist argument
2. the motivationalist argument

The Intentionalist

The intentionalist takes seriously the link between self-deception and intentional, interpersonal deception. Consequently, the intentionalist must attempt to overcome the static and dynamic paradoxes. Various at-

tempts have been put forward to rectify the two paradoxes, most of which include dividing the self in some fashion. An individual may be able to believe contradicting ideas as long as they only attentionally focus on one idea (Demos, 1960). Temporal partitioning permits an individual to actively seek to deceive a future self, thus avoiding ever simultaneously holding conflicting views (Bermúdez, 2000). Partitive intentionalism divides the mind into something analogous to the conscious and unconscious, suggesting that two parts can hold conflicting views (Davidson, 1987).

Intentionalists run into several problems, however. First, there is little to no empirical evidence demonstrating that simultaneously holding two beliefs occurs, regulating many intentionalists to a discussion of the hypothetical and imagined (Mele, 1997; Porcher, 2015). Second, the intentionalist may have to bite the proverbial bullet in accepting ambiguity in the definition of the self (Smith, 2014). Is the self composed of two or more homonculi who are attempting to deceive each other? If multiple mechanisms are deceiving each other within one person, how useful is it to discuss that person as having a single, homogenized identity (Kurzban, 2011)?

The Motivationalist

The motivationalist attempts to dodge the static and dynamic paradoxes by rejecting the notion that self-deception is inextricably linked to intentional deception. Mele (2001) suggests that self-deception need not require the intention to deceive, but rather, the self-deceived individual must be *motivated* to believe something which is false. Similarly, Bagnoli (2012) defines self-deception as "the acquisition and retention of a belief despite overwhelming evidence to the contrary." For example, two parents may be motivated to believe their child has not been drinking, despite the smell of alcohol on the adolescent's breath. Nelkin (2002) includes biases in information acquisition as potentially self-deceptive. For instance, the parents of the aforementioned child perform self-deception if they go out of their way to remain ignorant of the alcohol abuse.

Opponents argue that it is difficult for the motivationalist to differentiate between self-deception and simple wishful thinking (Deweese-Boyd, 2012; Smith, 2014). When is someone self-deceived and when are they just wishing the world was different? Smith (2014) worries that by disregarding the requirement of intention, we have also lost a firm understanding of self-deception's success criteria. A homogenized definition of the self has been saved, but potentially at the cost of expanding the idea of deception too broadly. Porcher (2012) contests that the motivationalist fails to disentangle the notion of both knowing and not knowing p. Why would an individual avoid information unless they are aware (on some level) that the information is likely to alter their belief? Thus, (on some level) the individual already knows that their current belief is false; consequently, they are not deceived. Lazar (1999) argues that the motivationalist view fails to account for instances of "twisted" self-deception (Mele, 1999) where the deceived would rather not have their current belief. For example, a jealous husband who is (incorrectly) convinced of his wife's infidelity may rummage for evidence, despite a desire to be wrong. Why would the husband be motivated to believe something he does not desire?

2.1.4 Where is a Functional Account?

Recently there has been a call for the inclusion of a functional explanation of self-deception. Baghramian and Nicholson (2013) suggest that examining how self-deception could survive natural selection may focus the philosophical discussion. The phenomenon may be better understood by considering when and how self-deception can provide an advantage. However, the inclusion of evolution into the philosophical conversation of self-deception remains surprisingly sparse. One exception is David Livingstone Smith's teleofunctional approach.

A Teleofunctional Approach

Smith (2014) attempts to unite the strengths of the intentional and motivational theories in what he deems a teleofunctional approach to self-deception. His goal is to retain the motivationalist's unified self whilst incorporating the intentionalist's ability to define self-deception's success criteria.

To Smith, an implicit criterion for self-deception is an ability to diagnose whether the act succeeds or fails. It is precisely this difficulty in diagnosing whether one has succeeded in self-deceiving that causes the motivationalist problems in differentiating between self-deception and wishful thinking (Smith, 2014). Smith (2014) attempts to circumvent this problem by calling upon the functional definition of deception (see Section 2.1.2) and arguing that self-deception must have a functional purpose.

Smith argues that the function of self-deception is the same as interpersonal deception: to induce false beliefs. Further, in order for the mechanism which induces false beliefs to be functional, it needs to have survived natural selection. Rather than getting lost in a debate about awareness and intention to deceive the self, Smith claims that humans could have evolved to deceive themselves if a mechanism which evolved to generate false beliefs inhibits another mechanism evolved to generate true beliefs. Just as inhibitory neurons have the function of limiting excitatory neurons, so might some process operate to dampen the functionality of another process whose function is to accurately represent the world. Formally, Smith defines self-deception as:

> O is self-deceived iff O possesses character C with purpose F of correctly representing some feature of its world, and character C* with purpose F* of causing C to misrepresent that feature, and it is in virtue of performing F* that C* causes C to misrepresent (Smith, 2014).

Smith bypasses the need for intentional deception by shifting to the requirement of evolutionary function. In doing so, Smith's definition can eloquently link self-deception with non-human definitions of deception. If one organism possesses a characteristic with the purpose of causing a different organism to misrepresent the world, then the first organism has deceived the other (Smith, 2014).

Smith makes significant strides in integrating a functional account into the philosophical discussion on self-deception. However, there may be several other ways the evolutionary process can clarify some of

the difficulties plaguing the definition of self-deception. Next, I discuss how a functional account of self-deception may help explain the dynamic paradox as well as twisted self-deception.

Can Evolution Help with the Dynamic Paradox?

Smith's use of the functional definition of deception may also aid in disentangling the dynamic paradox. The dynamic paradox questions how an individual could both intend to deceive and succeed at deceiving whilst being aware of the intention to deceive (Mele, 1997). Whilst, perhaps succeeding to deceive a self who is aware of the intended deception is impossible, a part of the self whose *function* is to deceive the self without awareness seems quite feasible. Certainly imitating flowers do not intend to deceive, but they do possess the function to deceive (Smith, 2004; Millikan, 1989). In this light, humans might function to deceive without the intention of doing so. The inclusion of a functional account of deception is another version of partitive intentionalism, but it is an important one which is rarely discussed.

The firm intentionalist may argue that an individual who evolved to be wrong is not self-deceived, but rather manipulated by evolutionary forces. If evolution, as opposed to the individual, is culpable for the deception, then the criteria for self-deception is not met. However, to me, the evolutionary discussion seems worth considering. Even if evolutionary forces are to blame, where does the force reside if not in the self? If the self is being deceived and the force performing the deception resides in the self, then is this not *ex vi termini* self-deception? Further, this notion escapes any risk of partitive intentionalism's competing homunculi; the individual does not intend to deceive, but is still deceived.

Can Evolution Help with Twisted Self-Deception?

In the case of twisted self-deception (Mele, 1999), an individual incorrectly believes something they hope is not true. In the canonical example, a wife believes on weak evidence that her romantic partner is unfaithful and scours for evidence, hoping she is wrong (Deweese-Boyd, 2012). The motivationalist has had difficulties integrating twisted self-deception into their model. Why would an individual be motivated to believe something they wish was not true (Lazar, 1999)?

Again, integrating evolution into the conversation may offer insight. The wife may be motivated to believe the falsehood due to naturally selected cognitive biases, even if she wishes she was not. Under the banner of Error Management Theory (further reviewed in Chapter 4), research has shown there are contexts where biasing decisions may prove optimal, even if unpalatable (Johnson et al., 2013; Haselton et al., 2015; Galperin and Haselton, 2012; Haselton and Buss, 2000; McKay et al., 2009). The intuition may be best explained by example. Imagine walking through a forest, attempting to differentiate between sticks and snakes. It may have been evolutionarily advantageous to bias one's beliefs to a snake, even though a stick is more likely (Galperin and Haselton, 2012). Being on guard against unlikely threats may have aided survival, even at the cost of heightened anxiety (Bateson et al., 2011).

The evolutionary costs of partner infidelity are significant, and may explain many of the present day

human mating behaviours (Buss, 2007). As a consequence, hyper-jealousy may serve an adaptive role, inhibiting partner infidelity (Schipper et al., 2006). This is supported by the fact that the contexts where individuals are most jealous differ by gender, and those contexts are correlated to the different gender's evolutionary risks (Easton et al., 2007). No one wants constant anxiety regarding snakes or infidelity; however, motivation can still exist without desire by considering selective pressures.

2.1.5 Summary

Clearly, the discussion of self-deception remains fraught with controversy. It seems the frailties of folk-psychology and the vernacular with which we discuss terms such as "belief" and "self" are brought to light in attempting to describe self-deception as a phenomenon (Porcher, 2015, 2012; Schwitzgebel, 2010). Further, the philosophical debate on self-deception has, in general, not included evolution (Baghramian and Nicholson, 2013). Smith (2014) offers some insight into how evolutionary mechanisms could deceive the self without the self's awareness. Additionally, I have argued that several of the apparent difficulties with self-deception can be explained when considered in an evolutionary framework. At the very least, including evolution into the philosophical debate seems warranted. With this in mind, and in hopes of clarifying some of the confusion, I turn to a review of the current theories regarding the evolution of self-deception.

2.2 Existing Theories for the Evolution of Self-Deception

At first glance, it seems nonsensical that an agent could effectively navigate the world when motivated to maintain false beliefs. As Trivers (2000) claims,

> For a solitary organism, the prospects [of evolving self-deception] seem difficult, if not hopeless. In trying to deal effectively with a complex, changing world, where is the benefit in misrepresenting reality to oneself?

What types of environments would offset this, generating selective pressure for misrepresentations? The research is still fairly nascent, only coming into its own in the past decade. In a recent review, Chance and Norton (2015) diagnosed three potentially adaptive functions of self-deception *a*) psychological benefits, *b*) interpersonal deception, and *c*) social status. Here, I review all three, but merge the benefits of social status under the banner of interpersonal deception, since, garnering social status through self-deception requires the deception of the other.

2.2.1 Defining the Evolution of Misbeliefs and Being Strategically Wrong

A note before I begin. I will also review literature on the evolution of misbeliefs as well as being strategically wrong. In the evolutionary theoretical literature, self-deception is not as meticulously defined as in the philosophical discourse. In fact, many do not distinguish the evolution of misbeliefs from the evolution of

self-deception. As such, I review the literature on each of these terms, in turn. However, if terms seem confused, it is probably because they are.

2.2.2 Psychological Benefits

Some argue that self-deception evolved to augment the psychological well-being of the self-deceived. Consider the example where the self-deceived parents continue to believe that their underage child has not been drinking despite the evidence. This archetypical example has been extensively employed by the motivationalist to suggest that by self-deceiving, the parent is able to maintain a desired psychological state. As such, self-deception can act as a psychological defence mechanism. By preserving the belief that their child is not using drugs, the parents can avoid a reality that might injure their psychological homoeostasis (Bortolotti and Mameli, 2012; Bagnoli, 2012).

Proponents: The Intrapersonal Benefits of Positive Illusions

Empirically, there is evidence to suggest that biasing information can lead to improved psychological well-being. On average, individuals tend to believe they are better than is objectively the case (Alicke, 1985; Alicke et al., 1995). This is true for a variety of tasks, such as driving (DeJoy, 1989; Dalziel and Job, 1997), avoiding divorce (Weinstein, 1980), and life expectancy (Weinstein, 1980). This optimism bias appears to be neurologically maintained; neuronal learning weights positive experiences higher than negative experiences (Sharot, 2011).

These positive illusions seem to confer psychological benefits. Exaggerated beliefs in one's intelligence is positively correlated to both life-satisfaction and self-esteem (Dufner et al., 2012). Those with higher self-esteem frame events of their life in such a way as to maintain their positive outlook (Brown, 2014). Cummins and Nistico (2002) argue that positive cognitive biases aid in maintaining life satisfaction and maintaining homeostatic well-being.

Depressive Realism is a theory coined by Alloy and Abramson (1979), who argue that mild to moderately depressed individuals hold a more accurate view of the world compared to controls. If accurately representing the world is correlated to depression, then positive illusions might aid in maintain a healthy well-being. A meta-analysis covering 75 experiments shows a significant effect for depressive realism (Moore and Fresco, 2012). Interestingly, in the meta-analysis, depressed individuals still demonstrate a positive illusion bias, just less compared to the control group. Related work shows that illusion of control is inversely correlated to depressive symptoms (Alloy and Clements, 1992).

Shelley Taylor is one of the most well-known proponents for the theory of adaptive positive illusions. Throughout her career, Taylor and her colleagues have found significant evidence to support the hypothesis that positive illusions lead to higher levels of self-esteem and mental health (Taylor and Brown, 1988). For instance, self-affirmation is positively correlated to reduced symptoms in early-stage breast cancer patients (Creswell et al., 2007). Higher levels of self-esteem and optimism are inversely correlated to

stress (Creswell et al., 2005). Importantly, the mental health benefits witnessed in the empirical literature cannot be explained by biases introduced by a small fraction of individuals with high levels of self-aggrandizement — the phenomenon is pervasive (Taylor et al., 1983).

Proponents: Psychological Defence Mechanisms

From the clinical perspective and hearkening back to early psycho-analysis, it has long been suggested that the mind is willing to confabulate in order to protect itself from the psychological devastation of reality (Freud, 1936, for a nice review see Vaillant, 1992). Butler (2000) considered a patient who suffered from a traumatic car accident, leaving him paralysed. A year after the accident, the patient developed Reverse Othello Syndrome; he began to believe that a former romantic partner and he had recently married. However, the former partner had left the patient soon after the accident. Extending the work of Jaspers et al. (1963), Butler (2000) argues that such delusions aid in maintaining intra-psychic coherence in the face of tragedy.

In fact, an assortment of clinical diagnoses appear to lend credence to the notion that the mind will protect itself with denial in the face of tragedy. Ramachandran (1996) considered a patient with anosognosia. Due to a right-hemispheric stroke, the patient was unable to walk or move one arm. However, she persistently denied her malady, claiming she could walk. Baird and McKay (2008) studied a patient with retrograde amnesia who self-aggrandized at a much higher rate than controls.

As previously mentioned, Ramachandran (1996) found that some right-hemispheric stroke victims deny their physical impairments. This led Ramachandran to define a theory for the neurological localization of self-deception. In the footsteps of Freud, Ramachandran and Blakeslee (1998) argue that the human mind evolved a mechanism to ignore information that threatens an individual's model of the world. As his theory goes, in a healthy mind this mechanism balances between when information should be ignored to maintain equilibrium, and when information calls for a reassessment of the mind's working model of the world. However, in patients suffering from anosognosia, this mechanism is broken and the psychological defence mechanism runs rampant, explaining the denial of pathology (for an excellent review of Ramachandran's theory see: McKay et al., 2005).

Proponents: Terror Management Theory

A variety of theories propose that persistent cognitive awareness of our mortality would have been detrimental to our ancestors' psychological well-being. Thus, it is adaptive, that we (generally) remain unaware of our impending mortality. Becker (1973) posits that if individuals recognized the frailty of their lives and the arbitrary nature at which events play out, they would have been stricken with anxiety and terror.

This led Greenberg et al. (1986) to suggest that self-esteem enhancing illusions may play an adaptive role in assuaging the pervasive anxiety which would follow an unbiased perception of the world. Under the name "Terror Management Theory", Greenberg et al. (1997) argue that evidence suggests that psychological well-

being is correlated to denial of reality. For example, Terror Management Theory predicts that individuals confronted with their eventual mortality will attempt to suppress the thoughts. It has been shown that after becoming aware of one's mortality, those under cognitive load demonstrate higher levels of death thought accessibility, lending credence to the hypothesis that suppression is the status quo (Arndt et al., 1997).

In a similar vein, Varki and Brower (2013) postulate that selective pressure for intelligence may have required self-deception. Particularly, they posit that whilst intelligence augments problem solving, it also unveils a comprehension of one's mortality. They suggest that the existential threat that comes from realistically processing the world may cost more than the benefits conferred by intelligence. It is only when intelligence coevolves with an ability to deny one's mortality (e.g. through optimism) that the benefits of intelligence outweigh the costs (Varki and Brower, 2013).

Proponents: Self-signalling

It has been argued that self-deception can arise through self-signalling (McKay et al., 2011b; Mijović-Prelec and Prelec, 2009). In ambiguous situations, the self may attempt to bias its beliefs to positive results. As Mijović-Prelec and Prelec (2009) say,

> Self-signalling implies that if the desire for good news is strong enough, it will bias interim assessments even if such biasing reduces overall chances of achieving the long-run goal.

Evidence for this theory is wide-ranging. Those primed with the knowledge that cold resistance is correlated to longevity will leave their hand submerged in cold water much longer than controls (Quattrone and Tversky, 1984). When given answers to an exam, students misattribute their high exam score to their own intellect (Chance et al., 2011). Even more, they do not learn to adjust this misattribution over multiple trials (Chance et al., 2015). Mijović-Prelec and Prelec (2009) demonstrated that, as financial incentives rise, individuals increasingly bias their decisions toward maintaining a positive outlook, even at the cost of financial loss.

Opponent: Self-Esteem is (at best) an Intermediate Step to Social Enhancement

There are, however, opponents to the idea that self-deception evolved to augment psychological well-being. von Hippel and Trivers (2011a) reject the notion, arguing that proponents conceptually misunderstand evolution. The evolution of any attribute must describe how the phenotypic behaviour associated with the attribute provides a selective advantage (Trivers, 2000). The hypothesis that self-deception evolved to enable feeling good about oneself fails to provide the argument for such an advantage (von Hippel and Trivers, 2011a; Kurzban, 2012; McKay et al., 2009). It would be analogous to suggesting that humans evolved to eat sugar because it made them feel better. Robert Kurzban bluntly agrees, "evolution doesn't care how happy you are" (Kurzban, 2012, p.136).

How would opponents explain the expansive experimental data demonstrating a positive correlation between self-esteem and positive illusions? Both von Hippel and Trivers (2011a) and Kurzban (2012) are proponents of the next theory I review — self-deception evolved to aid in interpersonal deception (see Section 2.2.3). As such, von Hippel and Trivers (2011a) argue that increased self-esteem mediates interpersonal deception. Self-esteem offers an evolutionary advantage in that it aids in deceiving the other. By feeling good about yourself, you can convince others that you are better than you are, thus tricking others into conferring undeserved social benefits.

Evidence seems to support this claim. Taylor et al. (1983) found that self-enhancement is positively correlated to both mental health as well as likeability. Further, self-esteem mediates likeability in self-enhancers (Dufner et al., 2012). Whilst augmenting self-esteem might not confer evolutionary benefits on its own, if it leads to increased social benefits, then self-enhancement may prove adaptive.

Opponent: Neurological Injuries May Not Elucidate Anything Other Than Pathology

The maintenance of false beliefs following neurological trauma does not necessarily lend credence to a functional explanation of self-deception (McKay et al., 2009). McKay et al. (2005) warn that we should be careful in differentiating between the maintenance of false beliefs as adaptive self-deception, and the maintenance of false beliefs because the neurological underpinnings for belief creation and acquisition are broken. For instance, anosognosia is only seen in victims with right-hemispheric strokes, those with left-hemispheric lesions acknowledge their malady (Ramachandran, 1996). If the denial of reality in the face of tragedy is adaptive, why don't left-hemispheric stroke victims suffer anosognosia? Further, Bisiach et al. (1991) found that spraying cold water into the ear of anosognosic patients eliminates the denial. What may first appear as adaptive self-deception may just be an example of the neural machinery gone awry. McKay et al. (2005) suggest that a framework for understanding delusions in clinical populations may be best served by amalgamating the motivational ideas of self-deception with the effects of damaging the neurological foundations of belief.

2.2.3 Trivers: Self-Deception Augments Interpersonal Deception

By far the most well-known theory on the evolution of self-deception comes from Robert Trivers (Trivers, 1991, 2011; von Hippel and Trivers, 2011a). Trivers (1991) hypothesizes that self-deception evolved to aid in interpersonal deception. His argument rests on the notion that while deceiving others can be advantageous, the cost of being detected is likely high. However, if a deceiving individual could reduce the probability of being caught, then deception might prove quite profitable, garnering the individual additional resources and increasing their evolutionary viability.

As a result, it seems plausible to expect selective pressure for improving deception in order to avoid detection. However, if individuals evolved improved deceptive capabilities, it seems just as likely there would be selective pressure for detecting deception (Cummins, 1999). Theoretically, this would result in

a runaway co-evolutionary Red Queen effect (Van Valen, 1973) between deception and detection — over generations individuals improve both deception and detection, but relative to each other little progress is made.

Trivers argues a way out of this co-evolutionary treadmill via self-deception. He proposes that the deceiving agent can circumvent the risk of detection by deceiving itself (Trivers, 1991). If an individual is unaware of their intention to deceive, then deception is difficult to detect. Further, even if they are detected, they maintain plausible deniability, reducing the costs of detection.

von Hippel and Trivers (2011a) attempt to unify most of the aforementioned experimental data under this theory. If self-enhancement, optimism, and overconfidence lead to improved social standing, then self-deception is advantageous, and thus may be adaptive. By believing one is better than is objectively the case, others are deceived and confer benefits to the self-deceived individual. While von Hippel and Trivers (2011a) reject the notion that self-deception could evolve to maintain psychological well-being (see Section 2.2.2), increasing psychological well-being may confer interpersonal advantages. For instance, while overconfidence increases self-esteem, self-esteem increases likeability (Dufner et al., 2012).

Self-Deception as Information Processing Biases

The definition of self-deception given by von Hippel and Trivers (2011a) is mostly in line with the motivationalist. They argue that self-deception is any information gathering/processing bias which favours "welcome over unwelcome information in a manner that reflects the individual's goals" (von Hippel and Trivers, 2011a, p.2). An individual can be self-deceived if they refuse to search for unwelcome information, or if they weight welcome information higher than unwelcome information. To von Hippel and Trivers (2011a) the self-deceived individual is motivated to be deceived. In this sense, they are in line with the motivationalist.

Supporting Evidence

There is evidence to support the hypothesis that self-enhancing positive illusions augments interpersonal deception by tricking others into bestowing unwarranted prestige. Overconfidence is linked to garnering higher social status (Anderson et al., 2012; Kennedy et al., 2013). Intellectual self-enhancing is correlated to likeability (Dufner et al., 2012). Self-enhancement is positively correlated with social attractiveness and social influence (Dufner et al., 2013). Self-perceived mate value is positively correlated to the popularity of the individual (Back et al., 2011).

Opponents: Deception is Untenable in Long Term Relationships

A plethora of opposition has been levied at this theory. Some suggest that evolving self-deception for interpersonal deception is unlikely since humans evolved in small groups. Surely the benefit of sharing

the truth amongst confidants outperforms deception[1] (Dunning, 2011). Further, plausible deniability only helps so much. Humans have long-term relationships and even unintentional errors eventually degrade a relationship (Bandura, 2011).

The experimental data shows mixed results regarding whether these interpersonal benefits outweigh the costs once the truth is discovered (Kennedy et al., 2013; Tenney et al., 2008). For instance, perceiving another's self-enhancement reduces social attractiveness, but boosts social influence (Dufner et al., 2013). Tenney et al. (2008) discovered that when overconfidence is verbalized, self-enhancing individuals are punished once the truth is known. In contrast, Kennedy et al. (2013) demonstrated that as long as communication is limited to non-verbal actions, the benefits of self-enhancement can be maintained even when the truth becomes known.

Opponents: People are Terrible Lie Detectors

Trivers' argument hinges on the idea that deception and detection of deception co-evolved. Because individuals are so successful at detecting deception, self-deception is given evolutionary traction. However, the empirical evidence on deception detection undermines this idea (Fridland, 2011; Vrij, 2011). People are terrible at lie detection, often perform hardly better than chance (Ekman and O'Sullivan, 1991). If individuals are poor at lie detection, then Trivers' argument falls apart. Self-deception is a high cost to pay for avoiding an unlikely event.

Recent evidence, however, suggests that poor deception detection may be the result of experimental artefacts. Individuals appear to perform better at deception detection in natural environments. Small groups of individuals are much better at detecting deception, compared to one lone individual (Klein and Epley, 2015). Further, historically deception detection was tested by offering participants facial clues, but a recent shift toward offering participants more contextual information has caused a significant jump in the likelihood of deception detection (Levine, 2015). If we are excellent at detecting deception in natural environments such as small groups saturated in contextual information, perhaps Trivers' hypothesis is feasible.

Opponents: Philosophical Problems

According to Smith (2011), von Hippel and Trivers (2011a) seem to be attempting to merge intentionalist and motivationalist ideas. On the one hand, they believe that biased information search is a sufficient condition for self-deception. As such, intentionalists contend that the argument presented by von Hippel and Trivers (2011a) is not self-deception. Biased information search is not self-deception, it is self-bias (Bandura, 2011; Pinker, 2011). Self-deception requires intention to deceive, and motivated information search fails to meet this criterion.

On the other hand, von Hippel and Trivers (2011a) attempt to retain intention in their definition of self-deception. Going so far as to argue that "self-deceivers are liars, whether they know it or not" (von Hippel

[1] I demonstrate in Chapter 3 that this is not true. Deception can actually evolve to *maintain* in-group cooperation.

and Trivers, 2011b, p.47). This does not sit well with those who believe deception requires a knowledge of the inaccuracy of the information (Vrij, 2011). If individual A really believes the proposition $\neg p$, even though the reality is p, then individual A is not deceiving individual B by relating $\neg p$; rather individual A is just wrong (Fridland, 2011).

von Hippel and Trivers (2011a) rebut this by suggesting that the intention to deceive can occur at the moment of information bias. Intentionally not seeking information so that you can deceive someone later is lying (von Hippel and Trivers, 2011b). However, I am not sure how this explanation coalesces with the above quote on how self-deceivers are liars, whether they know it or not. How can an individual lie without the intent to deceive, even if the intent is temporally adjusted?

Here, I will insert myself briefly into the discussion. As I mentioned in Section 2.1.4, a way out of this conundrum is to consider deception from a functional, evolutionary perspective. If individuals evolved a propensity to bias information search, then evolution is culpable for the deception. Whilst the individual may not intend to deceive, possessing traits which function to deceive is certainly closer to self-deception than just being "wrong". While von Hippel and Trivers (2011a) do not explicitly argue this, it seems implicit in much of Triver's work, particularly when defining the definition of regular, interpersonal deception (Trivers, 2011).

Opponents: Why not Consider the Evolutionary Advantages of Misbelief?

Having faced significant criticism regarding the definition of self-deception, some argue (myself included) that the best way forward is to focus on whether the information which elicits behaviour should always be veridical. In considering the argument by von Hippel and Trivers (2011a), Fridland (2011) says, "the real revelation would be to question the assumption that beliefs are most useful when true." Harnad (2011) agrees, positing "all that's needed for adaptive cognition and behaviour is information (i.e., data)." Kurzban (2012) argues that self-deception misses the mark. Since the self is modular and not unified, conversations about deceiving a unified self are incoherent (Kurzban and Athena Aktipis, 2007). Rather, Kurzban introduces the state of being *strategically wrong* — is it ever adaptive to act with incorrect information?

I agree with this criticism. Below, I argue that the use of false information in decision-making is a prerequisite for any evolutionary theory of self-deception (see Section 2.6). First, however, I turn to a review of the current literature on the evolution of misbeliefs and being strategically wrong.

2.2.4 The Evolution of Misbeliefs

To recall, I have defined misbelief and false belief interchangeably (see Section 2.1.1). Thus, this section focuses on the pressures which lead to adaptive false beliefs. Two points will become clear. First, the literature discussing the evolutionary feasibility of self-deception does not match the findings on the evolution of misbeliefs. This is despite the fact that both the self-deception and misbelief literature use very similar, if not identical, definitions of their phenomenon. Second, even within the literature on the evolution of misbeliefs

there is significant disagreement as to the evolutionary cause.

Heuristics / Ecological Rationality / Adaptive Rationality

Haselton et al. (2015) categorized the contexts where individuals "reliably produce representations that are systematically distorted compared to some aspect of objective reality." They demarcated three categories: *a*) heuristics, *b*) error management effects, and *c*) experimental artefacts (also see Haselton et al., 2009). I review the first two. Experimental artefacts are defined as instances where the experimental set-up is counter-intuitive to the evolved mind. As a consequence, individuals make suboptimal decisions because they would rarely (if ever) face such a dilemma in the real world. I skip a discussion of experimental artefacts, as they do not represent a context where misbeliefs are adaptive; rather, they represent situations where our evolved mind is vulnerable to systematic errors.

As Pinker (2005) points out, the mind is not necessarily adapted to uncovering the truth. Information is costly, requiring time and energy, "so a system designed for useful approximations (one that 'satisfices...') might outcompete a system designed for exact truth at any cost" (Pinker, 2005). For instance, it has been demonstrated that when seeking food, evolving a heuristic foraging rule outcompetes a rule seeking a perfect (but time consuming) understanding of the world (Higginson et al., 2015; Trimmer et al., 2008).

Cosmides and Tooby (1997) created a paradigm shift in how to frame suboptimal human decision-making. They argue that an understanding of human decision-making must be grounded in the historic context of our evolutionary past (Cosmides et al., 1992). Through a series of experiments, they argue that the human mind did not develop generalized intelligence, but rather adapted to meet specific evolutionary needs (Cosmides and Tooby, 2013). For example, individuals are much better at using the transitive property when it is framed in a social context, rather than framed in a mathematical context (Cosmides, 1989). Change-blindness (when an individual does not recognize the difference between two images) is significantly reduced when the images of animals are altered compared to inanimate objects, lending credence to the hypothesis that humans evolved to prioritize attention to those with agency (New et al., 2007). Humans still assess another's strength by facial expressions linked to barring teeth, much like our primate cousins (Sell et al., 2014). This is related to the idea of bounded rationality, that humans attempt to make the best decisions given the information and cognitive limitations they possess (Simon, 1972, 1991).

Noisy, Autocorrelated Environments

Fawcett et al. (2014) reviewed the literature on the evolution of cognitive biases. They argue that several biases can be attributed to noisy environments with autocorrelation. For example, after a series of bad experiences, clinical levels of depression and the associated lethargy is adaptive in autocorrelated environments (Trimmer et al., 2015). Further, the hot-hand fallacy is adaptive given autocorrelated environments (Wilke and Barrett, 2009). Predicting that a behaviour will continue is a valuable heuristic.

Several cognitive biases are also beneficial in noisy environments. When information is obscured, bi-

asing knowledge of your own strength can be adaptive (Ramirez and Marshall, 2015; Johnson and Fowler, 2011). When healing is costly, the placebo effect is adaptive in noisy environments (Trimmer et al., 2013). When the safety of an environment is hidden paying the cost to repair may not be advantageous. If others (e.g. doctors) clarify that an environment is safe, then the energy required for healing can be spent.

Error Management Effects

Error Management Theory (EMT) predicts that in noisy environments, where different actions lead to asymmetrical benefits/costs, biasing information when making a decision can be optimal (Johnson et al., 2013; Buss and Kenrick, 1998). The underlying idea is well explained by an example from McKay et al. (2009). If building a perfect smoke detector is impossible, it is better to error on the side of caution. Sounding an alarm when there is no fire is annoying, but not sounding the alarm when there is a fire is deadly. Thus, it is advantageous to bias the detector toward hyper-sensitivity. Even though this results in more errors, it results in paying a lower cost — less death, but more annoyance.

EMT has predominately been used to argue that, in unpredictable environments, it is advantageous to bias one's belief in the probability that an event will occur. As previously mentioned, humans tend to optimistically bias their predictions about the likelihood of a positive experience (Sharot et al., 2007). EMT demonstrates that in an unpredictable world such biases can be optimal by compensating for noisy environments (Haselton and Nettle, 2006; McKay et al., 2009; Sharot, 2011). Further, EMT has been used to describe how biased probabilistic thinking may explain human behaviour related to anxiety (Bateson et al., 2011), belief in god (Johnson, 2009), perceived sexual attraction (Haselton and Buss, 2000), and social ostracism of potentially diseased individuals (Kurzban and Leary, 2001). In Chapter 4, I show that EMT can also explain why individuals consistently mispredict how they will feel after an event.

The Evolution of Being Strategically Wrong: Interpersonal Deception

Kurzban (2012) uses slightly different vernacular compared to "misbelief". He is interested in diagnosing when it is beneficial for an individual to be *strategically wrong*. Despite the shift in language, Kurzban is essentially discussing the evolution of misbelief. "Wrongness" is identical to a false belief. Further, to Kurzban, "strategic" means that being wrong confers some advantage. In this sense, we can group adaptive misbelief and being strategically wrong, together. Both require that a false belief confers some benefit relative to veridical beliefs.

As previously mentioned, Kurzban (2011) believes the discussion of self-deception is confounded. If, as he purports, the mind is comprised of several components each tasked with different job (Kurzban and Athena Aktipis, 2007), then the notion of a unified "self" which is deceived just needlessly confuses the discussion. However, Kurzban (2012) agrees with Trivers, in that interpersonal deception is the only viable avenue for being strategically wrong; he rejects the notion that the self can benefit from a lack of veridical information, unless another party is also deceived.

Most Biased Behaviour Does not Require False Beliefs

McKay et al. (2009) argue that many of the aforementioned biases do not require a belief. As such, they should not be grouped under the heading of misbelief. They suggest that misbeliefs are only adaptive in the case of positive illusions. While error management tasks can lead to biased behaviours, they do not necessarily lead to biased beliefs (McKay and Efferson, 2010). For instance, individuals look both ways before crossing the street, even when they are fairly certain there is no oncoming traffic. Such individuals do not have a false belief that traffic is present, they realize that the high cost of a mistake is worth double-checking (McKay et al., 2009). Mckay and Dennett argue that positive illusions are the only misbelief where the falseness of the *belief* is a requisite for the advantage gained.

Opponent: Alter Decision Rules, not Information

Pinker (2011) argues that when designing a system, it is always better to bias at the decision-making level, rather than at the perceptual level. Imagine you put on a sweater when the thermostat reads less than 60°F. However, you discover that you are cold when it is 62°. To solve this problem, you can either *a*) change your decision rule so that you put on a sweater when it is 62°, or *b*) alter the thermostat so that it displays two degrees less than the actual temperature. Pinker argues it is always better to change one's decision rule rather than manipulating the information one receives; otherwise all systems which use the thermostat will receive biased information (Pinker, 2011).

Related, Trimmer et al. (2011) analysed whether it is advantageous to add bias when attempting to learn the probability of an event occurring. They showed that adding a bias is advantageous depending on the learning algorithm exercised. Given a reasonable prior, Bayesian updating does not require a bias (Marshall et al., 2013b). However, it is well accepted that human minds cannot deal with the complexities of Bayesian updating. Consequentially, given the cognitive limitations of comprehending the probability of success, approximating Bayesian rationality through heuristics may involve biasing one's beliefs (Marshall et al., 2013a; McKay and Efferson, 2010).

2.3 Summary: Three Roadblocks to the Theoretical Study of Self-Deception

Having reviewed the current literature regarding the evolutionary feasibility of self-deception, misbelief, and being strategically wrong, I believe further research is currently limited by three factors. First, within the theoretical literature, the definition of self-deception, and the related concepts of misbelief and being strategically wrong, are mired in confusion. Second, despite significant research, there is little consensus on whether any given theory succeeds in explaining functional self-deception. Finally, to my knowledge, no one has proposed a systematic framework for analysing the types of environments which could lead to the

evolution of self-deception. I discuss each of these points in turn.

2.3.1 What Are We Talking About, Again?

The philosophical and evolutionary theoretic conversations on the deceptive aspect of self-deception are disjointed (Mitchell, 1986). Whilst philosophers have focused on discussing the need for deception to be intentional, many theorists only require a false belief. For the theorists who believe self-deception provides psychological benefits, it seems the only requirement is an evolved process which generates self-enhancing illusions (Taylor and Brown, 1988; Ramachandran, 1996; Greenberg et al., 1997; Varki and Brower, 2013). For those who argue self-deception evolved to aid in interpersonal deception, the only requisite is an evolved mechanism for generating self-enhancing errors or biases to information gathering (von Hippel, 2015; von Hippel and Trivers, 2011a). Neither hypothesized solution requires any conscious intention or motivation on the part of the self-deceived. It seems that much of the research on self-deception and misbeliefs have been focusing on a similar, if not an identical phenomenon. The only requisite is an adaptive false belief.

2.3.2 Significant Disagreement Remains

If work on the evolution of self-deception and misbelief have been focusing on the same phenomenon, why are the proposed solutions so disparate? The potential selective pressures for the evolution of self-deception are still hotly debated. Whilst some argue that self-deception increases psychological well-being (Taylor and Brown, 1988; Ramachandran, 1996), both the empirical results and evolutionary coherence of the theory are contested. Not only are the clinical examples potentially miscategorized (McKay et al., 2005), the entire hypothesis may confound the requisites of natural selection (von Hippel and Trivers, 2011a; Kurzban, 2012; McKay et al., 2009). Opponents to the idea that self-deception evolved for interpersonal deception argue that such behaviour is untenable given the small, close-knit groups in which humans evolved (Dunning, 2011; Bandura, 2011). Further, even if the theory is sound, the presumption that evolution favoured deception detection mechanisms is controversial, particularly given how poorly individuals detect liars (Fridland, 2011; Vrij, 2011).

Kurzban (2012) agrees with the idea that interpersonal deception causes selective pressure for false beliefs. However, he and others argue that self-deception is a confounded term, which should be replaced with a study of the evolution of false beliefs (Harnad, 2011; Kurzban and Athena Aktipis, 2007; Fridland, 2011). McKay (2011) agrees with Kurzban's argument regarding interpersonal deception, but believes it is too limited in scope, suggesting that positive illusions should be included. Galperin and Haselton (2012) argue against the conclusions that positive illusions are the only existing selective pressure for false beliefs (McKay et al., 2009). While Error Management Theory does not require a cognitive bias to impart behavioural action, humans seem to exhibit the cognitive bias nonetheless.

2.3.3 *Post hoc* Categorization

Even if we ignore the lack of cohesion surrounding the research on self-deception, another problem exists. There has been no systematic analysis diagnosing the contexts where self-deception can evolve. Most of the discussion has focused on analysing one theory over another. To my knowledge, no one has classified the *types* of pressures which would lead to the evolution of self-deception. The few works which have focused on attempting to categorize the selective pressures leading to self-deception have done so via a post-hoc analysis. They considered all of the current research on the evolution of cognitive biases and classified them into groups (Haselton et al., 2015; Fawcett et al., 2014). This has been quite helpful, but it does not diagnose any way of searching for additional contexts where self-deception is adaptive.

In the next section, I attempt just such a diagnosis. I suggest the research on the evolution of self-deception can be aided by an analysis of the conditions required for the evolution of the phenomenon. By analysing the evolution of self-deception's constituent parts avenues of research may be discovered, shedding light onto the adaptive nature of the phenomenon as a whole.

2.4 Requisites for the Evolution of Self-Deception

Here I attempt to unify the philosophical definitions of self-deception with the theoretical discussion, and converge onto the necessary conditions required for the evolution of self-deception. I argue that almost everyone (intentionalists, motivationalists, and assorted evolutionary theorists) agrees that adaptive false belief is a necessary condition for the evolution of self-deception. Whilst some may believe it is a sufficient condition for self-deception, if almost everyone believes it is a necessary condition, then further research into the evolutionary viability of false beliefs should lend insight into the evolvability of self-deception.

2.4.1 Necessary Condition 1 : The Self-Deceived's Behaviour Must Provide an Advantage (or At Least Not a Disadvantage)

As mentioned in Chapter 1, one of the basic tenets of evolution, is that "natural selection would inevitably tend to preserve those individuals which were born with constitutions best adapted to any country which they inhabited" (Darwin, 1872, p. 161). A theory on the evolution of self-deception receives no exception to this criterion (Trivers, 2000). If self-deception were to evolve, we would expect that, in some environments, the behaviour of the self-deceived individual would perform at least as well as the undeceived other. If the self-deceived individual outperformed the undeceived, then the behaviour would likely be selected for, since the deceived individual would gain a reproductive edge. If the self-deceived performed as well as those with veridical knowledge, then self-deception could neutrally drift into the population. As long as self-deception is not detrimental relative to others, then it could drift into a population via its innocuous nature[2].

[2]In Chapter 7, I demonstrate how this can explain why so many individuals employ suboptimal strategies when playing certain economic games.

Importantly, the aforementioned philosophical definitions are not constrained by the need to outperform those with veridical knowledge. To both the intentionalist and the motivationalist, self-deception can be (and in fact more often than not is) detrimental. However, for self-deception to survive natural selection, it would need to lead to behaviour which provides an advantage relative to others.

2.4.2 Necessary Condition 2: The Self-Deceived Possesses a False Belief

All definitions of self-deception require that the self-deceived individual possesses a false belief. As Mele (2012) puts it, if an individual has a true belief, how are they deceived? Even if someone tricked them into a true belief, they still hold the true belief. They may have been deceived *into* believing a truth, but they are not deceived. David Livingstone Smith agrees, noting that "what intentional and non-intentional varieties of deception have in common is their having the purpose of inducing states that I have characterized as misbelief" (Smith, 2014, p.189).

2.4.3 Necessary Condition 3: The False Belief Must Promote Adaptive Behaviour

Natural selection only acts on the phenotype. So, when can natural selection act on a predisposition to possess a false belief? Historically, evolutionary theorists have differentiated between a belief and the behaviour the belief engenders (Cloak, 1975; Dawkins, 1982; Dennett, 1995). Possessing a false belief does not require that the belief is expressed through behaviour. For instance, an individual could deceive themselves into believing that the moon is made of cheese, but if they never use that information whilst making a decision, they cannot be selected for or against compared to others.

For self-deception to evolve, however, it must confer some benefit (Condition 1). If a necessary condition of self-deception is the possession of a false belief (Condition 2), then the last necessary condition unites the first two. The force driving the adaptive behaviour (Condition 1) must be the false belief (Condition 2). Importantly, as in the case of Condition 1, this is not a necessary condition for self-deception, but rather the evolution of self-deception.

2.4.4 Motivationalism Falls Out

If the three conditions are met, motivationalism appears to fall out. To the motivationalist, self-deception not only requires a false-belief, but a motivation to maintain that belief. Evolution provides that motivation. If employing a false belief generates a selective advantage, then there is reason to maintain it.

2.4.5 Intentionalist Requirements

These three conditions seem to represent necessary but insufficient criteria for the intentionalist. Condition 1 is an evolutionary requirement, and thus irrelevant to the strand of self-deception to which one adheres. Condition 2 is easily met, as the intentionalist acknowledges the difficulties in resolving the static paradox

— where someone both possesses a true and a false belief. While Condition 3 is not a requisite for the intentionalist, it is, again, a requisite for the evolution of self-deception.

The intentionalist, however, must also resolve the dynamic paradox, whereby the self-deceived individual must intend to deceive his/herself. Section 2.1.4 discusses how introducing an evolutionary framework may help resolve this dynamic paradox. However, even if the intentionalist does not support slipping into a functional definition of self-deception, the hard-line intentionalist still requires the three aforementioned conditions.

2.4.6 Theorists' Requirements

As mentioned in Section 2.3.1, the evolutionary theorists mostly define self-deception as the maintenance of a false belief. Whether self-deception improves one's psychological well-being, or aids in interpersonal deception, the only requisite is an adaptive misbelief. Consequently, I would be surprised if most of the aforementioned theorists were not willing to accept the evolution of false beliefs as a sufficient (let alone necessary) condition for the evolution of self-deception.

2.4.7 Requirements of Misbelief and Being Strategically Wrong

These three conditions are sufficient for the evolution of misbelief and being strategically wrong. Condition 2 requires the possession of a misbelief; by definition this is a requirement for the evolution of misbelief and being strategically wrong. The only other requirements are evolutionary constraints, namely that the misbelief provides an advantage compared to agents with veridical beliefs.

2.4.8 Summary

In summary, despite continued debate over the definition and aetiology of self-deception, almost all definitions require that an individual holds a false belief (Condition 2). Further, for the phenomenon to survive natural selection, the false belief must be converted into a behaviour (Condition 3), and that behaviour must provide an advantage compared to others (Condition 1).

The motivationalist and intentionalist may not agree that the inclusion of a functionalist approach to self-deception adds coherence to the debate. The motivationalist may argue that the motivation must come from the conscious awareness of the individual, and not from evolutionary forces. The intentionalist may suggest that shifting culpability of the intention to deceive onto evolution removes all the interest in the phenomenon. However, I suggest that even they must concede that a minimum requirement for the evolution of self-deception is that a false belief provides some advantage.

If the motivationalist requires conscious awareness, then the requirement of an adaptive false belief is not a sufficient condition, but it is still a necessary one. Similarly with the intentionalist; a conscious requirement for intention to deceive does not remove the necessary requirement of the false belief. Further, much of the

theory regarding the evolution of self-deception, misbelief, and being strategically wrong all focus on the evolution of misbelief. In summary, the need for the evolution of a false belief is foundational to the concept of self-deception, regardless the nuances of the definition to which one ascribes.

2.5 The Difficulty in Analysing the Evolvability of Beliefs

If theory on the evolution of self-deception can be extended by diagnosing contexts where false beliefs are adaptive, a large problem still looms. How does one theoretically analyse the evolvability of a false belief? Possessing a false belief does not necessarily mean it is employed in decision-making. For instance, one can imagine an agent who believes $\neg p$ but employs p during action selection.

The concept of "belief" has a long philosophical history. The extent to which a belief must be tied to behaviour is the battleground upon which much of the discussion has been held (Schwitzgebel, 2015). Beliefs may or may not be linked to behaviour, or may not even exist other than as a useful folk-psychology construct (Churchland, 1981). Further, however one defines "belief", perhaps the only way to diagnose the existence of a belief is through behaviour (Dennett, 1987). It seems, however, that, at least occasionally, beliefs and behaviour misalign (Gendler, 2008b). The term *alief* has been used to explain instances where one's beliefs and behaviour are disjointed (Gendler, 2008a). For example, when an individual shakes uncontrollably, despite believing they are safely behind a railing at the Grand Canyon.

If the evolutionary viability of self-deception rests on misbeliefs contributing to an adaptive behaviour (Condition 3), and if we cannot definitively tie belief to behaviour, then how can theory move forward? I suggest this issue can be sidestepped by discussing the utility of acting on false information.

2.6 Moving Forward: The Evolution of Employing False Information

We can bypass the debate regarding the evolvability and definition of beliefs by considering whether contexts exist where employing false information in decisions outperforms those with veridical information. For false beliefs to evolve, they must generate adaptive behaviour (Condition 3). As such, the belief must be used in the behaviour. This eliminates the aforementioned quandary where an agent believes $\neg p$ but employs p during decision making. The false belief of $\neg p$ must be used in generating behaviour. Thus, the evolution of false beliefs requires the use of false information.

Evolving the use of false information does not, of necessity, require the existence of a false belief. However; the evolution of false beliefs require the use of the false information. This leads to an important inference:

> *If all definitions of the evolution of self-deception require the evolution of false beliefs, and the evolution of false beliefs requires behaviour dictated by false information, then the theory of*

self-deception may be expanded by diagnosing environments where it is adaptive to use false information.

Assuming this inference and the three necessary conditions for the evolution of self-deception are valid, then research on the theoretical feasibility of the evolution of self-deception may be extended by discovering environments where employing false information provides an advantage over those utilizing veridical knowledge. This dissertation attempts to do just that; extend the knowledge of environments where it is adaptive to bias or ignore veridical information.

If one accepts that false beliefs are sufficient for self-deception and that false beliefs are often used in behaviour, then this dissertation argues at the ubiquity of situations where self-deception is adaptive. If one accepts these conditions as necessary but insufficient for the evolution of self-deception, then this work still offers much needed insight. It demonstrates that, at minimum, one of the necessary requirements for the evolution of self-deception is ubiquitously found. Either way, the evolutionary feasibility of a necessary condition is a good place to begin a systematic search of the phenomenon as a whole.

2.7 Conclusion

In this chapter, I reviewed the state of the literature on both the definition and evolutionary viability of self-deception. Debate continues on both fronts. Intentionalists argue that the self-deceived must intend to deceive themselves. Motivationalists suggest that biased information search is sufficient. The hypotheses surrounding the evolution of self-deception are varied and not particularly focused, ranging from demanding interpersonal deception, to augmenting self-esteem. Little attention, however, has been paid to a systematic search for the types of environments which select for self-deception. No one appears to have diagnosed the necessary conditions for the evolution of the phenomenon, let alone carried out the search for the feasibility of those conditions.

Here, I have set forward an argument for the necessary conditions for the evolution of self-deception. Regardless the definition of self-deception (of which there are many), its evolutionary viability depends on generating adaptive behaviour through false information. As such, self-deception may benefit from a more rigorous investigation into environments which generate selective pressure for lacking veridical information.

For the remainder of this dissertation I begin just such a search. Over the next six chapters I prove that there are a variety of previously undiagnosed contexts where biasing or ignoring veridical information outperforms those who employ veridical information. I resolve several open scientific questions by demonstrating that it is adaptive to lack veridical knowledge. Further, this work may begin to explain the ubiquity of self-deception. Since selective pressure for lacking veridical knowledge is one of the necessary conditions for the evolution of self-deception, then it may be time to take serious the idea that self-deception is playing a role in human evolution.

3

Value Homophily Benefits Cooperation but Motivates Employing Incorrect Social Information

> *Once in a while, though, he could not help seeing how shallow, fickle, and meaningless all human aspirations are, and how emptily our real impulses contrast with those pompous ideals we profess to hold. Then he would have recourse to the polite laughter they had taught him to use against the extravagance and artificiality of dreams; for he saw that the daily life of our world is every inch as extravagant and artificial, and far less worthy of respect because of its poverty in beauty and its silly reluctance to admit its own lack of reason and purpose.*

H.P. Lovecraft, *The Silver Key*

> *The most astonishing thing of all, about man's fictions, is not that they have from prehistoric times hung like a flimsy canopy over his social world, but that he should have come to discover them at all.*

Ernest Becker, *The Fragile Fiction*

3.1 Summary

Individuals often judge others based on third-party gossip, rather than their own experience, despite the fact that gossip is error-prone. Rather than judging others on their merits, even when such knowledge is free, we judge based on the opinions of third parties. Here we seek to understand this observation in the context of the evolution of cooperation. If individuals are being judged on noisy social reputations rather than on certain merit, then agents might exploit this, eroding the sustainability of cooperation. We employ a version of the prisoner's dilemma (the donation game) which has been used to simulate the evolution of cooperation through indirect reciprocity. First, we validate the proposition that adding homophily (the propensity to interact with others of similar beliefs) into a society, increases the sustainability of cooperation. However, this creates an evolutionary conflict between the accurate signalling of in-group status versus the veridical report of the behaviour of other agents. We find that conditions exist where signalling in-group status outweighs honesty as the best method to ultimately spread cooperation.

3.2 Introduction

It is frequently argued that the key advantage which drives the evolution of social learning compared to individual learning is that it provides more or better information at a lower cost. An individual that can benefit from what others know can draw knowledge from a wider range of experience at lower personal risk than one limited to their own immediate life events (Boyd and Richerson, 1985; Fernández-Juricic and Kacelnik, 2004; King and Cowlishaw, 2007; Magurran and Higham, 1988; Rendell et al., 2010). However, what happens when an individual discovers that the socially-received information is false? If correctness is the paramount concern, we might expect that false socially-learned information would be replaced by a more reliable source (e.g. a first-hand experience).

However, there is mounting evidence that humans do not do this. Sommerfeld et al. (2008) tested the circumstances under which a participant would donate money to another individual. In each round, participants were paired and one person (the donor) was offered the opportunity to donate to the other (the recipient). Each donor was given either: *a*) the directly-observed history of the receiver's tendency to donate when the receiver had been a donor; or, *b*) the gossip-spread reputation of the receiver from third parties. Significantly more variation in the tendency to donate was explained by individuals use of reputation compared to their use of direct observation. Furthermore, Lorenz et al. (2011) showed that individuals edit their answers to questions based on other people's responses, though this often makes the average response of the group less correct. Even compared to other species, humans continue to persist with inaccurate social views longer than other primates (Whiten et al., 2009).

The null hypothesis is that the above behaviours are maladaptive exceptions to what is typically an adaptive heuristic. Social learning could be the best strategy despite a high incidence of error when the full cost of accruing accurate information, including time, is taken into account (Mitchell et al., 2013; Bryson

et al., prep). Further, researchers have proposed multiple heuristics by which humans bias their search for the most useful socially-acquired information. Conformity bias — acting with the majority (Henrich and Boyd, 1998), prestige bias — imitating the most prestigious (Henrich and Gil-White, 2001), pay-off bias — imitating the most successful (Mesoudi, 2011) are examples. Additionally, although social information transmission may introduce error, so may individual learning. Thus, in the rare situations where correct direct observation is easily attainable (e.g. Sommerfeld et al. 2008), individuals may employ noisy social information instead of correct directly observed information, because typically direct observation is expensive or similarly error prone.

These explanations argue for error prone social learning as the 'least-worst' option, and that the human tendency to employ social information in contexts where it is not useful is merely a local exception to a generally adaptive heuristic. However, the underlying assumption is that the utility of information (whether gossip or asocial) rests upon the accuracy of the information. Here we propose an alternative explanation for ignoring accurate personal experience in favour of social information. If social information comes with social prescriptions as to the employment of that information, then the factors influencing one's decision to utilize the information may extend beyond accuracy alone.

We demonstrate that ignoring veridical personal experience, in a particular context, can facilitate the cooperative exchange of information more generally. In particular, the mechanisms that generally facilitate cooperation can create a dilemma between two levels of information: *a*) information about the transmitter, and *b*) information to be transmitted. We begin with a model of society where cooperation is regulated via reputation. Agents decide whether to donate to other agents and the reputation of the agent is spread throughout the population. We show that when homophily (the tendency to act with others who share similar beliefs) is added to this model, the robustness of cooperation is increased against error in communication. However, as a consequence, it becomes adaptive to employ incorrect social information even when an individual agent has access to correct information. In conditions where the pay-offs for group unanimity outweigh the costs of acting based on inaccurate information, there is selective pressure for norm-following.

Our examination employs both computer simulations and formal analysis and proceeds as follows. First, we briefly introduce the literature on homophily and the evolution of cooperation. Next, we model the donation game to examine the effects of error on the evolution of cooperation. The donation game has been utilized as an existence proof for the evolution of cooperation in highly mobile societies (Nowak and Sigmund, 2005). It can be described as a specific instantiation of the prisoner's dilemma (Suzuki and Kimura, 2013; Masuda, 2012; Uchida and Sigmund, 2010) and continues to be used for studying cooperation both theoretically (Tanabe et al., 2013; Masuda, 2012; Hilbe et al., 2013; Stewart and Plotkin, 2013; Uchida and Sasaki, 2013; Marshall, 2011, 2009; Nakamura and Masuda, 2012) as well as experimentally (Angerer et al., 2014; Sommerfeld et al., 2008). We show that the donation game and the spreading of reputation can be used to sustain cooperation (Panchanathan and Boyd, 2003; Nowak and Sigmund, 2005). The result is employed as a baseline for measuring cooperation.

Next, we analyse the effects of value homophily (the propensity to interact with those who share your

beliefs) on cooperation. We find that as interactions become biased toward shared beliefs, cooperation becomes increasingly robust to error. Finally, we allow individuals to discover in isolation whether the social information they have received is incorrect. We test the consequences of acting on this information. We find that in homophilous societies, agents employing correct information are invaded by agents communicating known error. This demonstrates that honest signalling about own-group membership can outweigh the importance of honest signalling about others behaviour. We discuss some of the consequences of the results for the literature on self-deception.

3.3 Model & Context

3.3.1 The Problem of Cooperation

In order to explore these issues, we need a context which meets certain requirements. First, the agents must learn valuable information socially. Second, that information must be subject to error. And finally, individuals must possess the ability to overrule what they socially learn, but in doing so breach a social norm. For our model, we implement a version of the Prisoner's dilemma, called the donation game (Marshall, 2011). This game has been used to show that cooperation can be established in a society when individuals exchange social information about the reputations of others (Nowak and Sigmund, 1998, 2005). We will give more details of the model in the following section, but first we will give some background information on the problem of cooperation.

A cooperative society is defined as one in which individuals benefit from the collective absence of defection (Axelrod and Hamilton, 1981). However, it is often the case that for any individual member, defection is advantageous when others are cooperative. Several mechanisms have been hypothesised to overcome this problem of defection, notably reciprocal altruism (Trivers, 1971). In reciprocal altruism, an agent behaves prosocially with another so that the other will reciprocate at some later date. However, mobile societies, such as human ones, are often seen as vulnerable to free-riders (Enquist and Leimar (1993) though see Schonmann et al. (2013)). Individuals might defect opportunistically and move on before the consequences of their behaviour can catch up with them. In such cases, a different mechanism may be required to explain cooperation.

Indirect reciprocity (Nowak and Sigmund, 1998) solves this problem as an agent behaves prosocially with another because it is likely to subsequently receive a benefit from a different agent. This can be achieved when individuals observe each other, judge behaviour according to a norm, and pass on the resulting reputation via social transmission. Defectors can no longer free-ride, however mobile they are, so long as for every interaction they are likely to be preceded by their reputation and suffer a cost.

It should be clear that accuracy of information can be measured: for example, how closely an individual's reputation matches their actual behaviour. But to test the hypothesis described above, we need information to have further normative implications. We therefore need to distinguish between first and second order

Action	Reputation
Cooperate	G
Defect	B

Table 3.1: First order norm. An agent receives a good reputation for performing the cooperate action, regardless of the reputation of the recipient. An agent receives a bad reputation for not cooperating.

	Good	Bad
Cooperate	G	B
Defect	B	G

Table 3.2: Judging norm. An agent receives a good reputation for cooperating with good, or defecting against bad agents. This second order norm can enforce cooperation but creates scope for descriptive/normative conflict. See further the main text.

judgements.

3.3.2 First and Second Order Norms.

Previous studies on how the spread of reputation can establish cooperation employed a first order norm (see Table 3.1). If an agent cooperates, it receives a good reputation. Reputationally-aware agents cooperate with good agents, or defect against bad. But employing first-order norms introduces a threat to a cooperative society from an unexpected source: those who do good indiscriminately (Nowak and Sigmund, 1998). In a cooperative society, indiscriminate cooperators may invade the population via neutral drift. Whilst indiscriminate cooperators do not directly introduce defection, since they do not pay heed to the reputation of others, a society comprised of indiscriminate cooperators is unable to punish defectors via exclusion, and thus is likely to be invaded by them. Therefore, it is not sufficient for cooperation to exploit a reputation for doing good or bad — a first order judgement. Instead, cooperation also requires reputational impact for doing good to the good and bad to the bad — a second order judgement (see Table 3.2).

However, the move to a second order norm cannot be made without introducing the risk of conflict between actions motivated by social versus asocial knowledge. When an individual interacts with another who is socially negatively reputed, but discovers in isolation that this partner's reputation is false, what is the individual's best strategy? If they respond 'honestly,' that is in accordance with their descriptive knowledge about their partner, the actor risks putting itself in breach of the norm their peers employ.

3.3.3 Descriptive versus Normative

We term as *descriptive* aspects of knowledge where the utility depends on its accuracy with respect to the target of its description. Our hypothesis, tested here, is that there may be instances where the utility of the information is not solely constituted by its accuracy. To illustrate how social information may come with implications beyond the solely descriptive, consider the following thought experiment. One friend advises another against a certain restaurant, but despite the advice the latter goes there to dine. We ask the question: regardless of the eventual quality of the restaurant, does being out of phase with one's peer have implications?

When an individual accepts information from others, the hypothesised aspect of knowledge which goes beyond the descriptive, we term *normative*. The distinction between normative and descriptive becomes salient when the individual possesses accurate personal experience which contradicts the socially-held view. In this case, if the utility of the information is solely descriptive the best strategy is obvious: the individual benefits from overruling the inferior source of information with the more reliable one. But if the social information comes with implications beyond direct evaluation, the choice is not so simple. The benefit gained from prioritizing accuracy of information may be outweighed by the costs incurred in terms of reputation. In this case, individuals may gain by employing information they know to be false if they prioritize a normative response to gossip over a purely descriptive one. In this paper we explore the consequences of agents making such a choice.

3.3.4 Value Homophily

The likelihood of being faced with this dilemma not only depends on the amount of error in gossip, but also the topology of the social network on which the gossip transmits. Various studies have considered the implications of social network structure on cooperation (Ohtsuki et al., 2006; Santos et al., 2008; Taylor et al., 2007). However, we explore the consequences of varying the probability of an agent interacting with another who shares the same socially inaccurate information. To do this, we use the concept of value homophily.

Value homophily is the tendency to associate with individuals who share similar beliefs and values (McPherson et al., 2001). Computational models have argued the evolutionary feasibility of homophilous behaviour biases (Fu et al., 2012) and that the propensity for homophily can lead to local cultural convergence, with disparate global beliefs (Axelrod, 1997).

Furthermore, value homophily has been prevalently diagnosed within human society. Curry and Dunbar (2013) showed that shared hobbies, moral beliefs, and a sense of humour are correlated to frequency of communication between friends. Humans more frequently interact with others who share values in dating (Fiore and Donath, 2005), drug use (Kandel, 1978), and several other self-reported personality traits (Adamic et al., 2003). Humans expect like-minded individuals to be more intelligent, moral, and knowledgeable of current events (Byrne, 1961). Ross et al. (2013) studied the differences in a folk-tale across geographic and

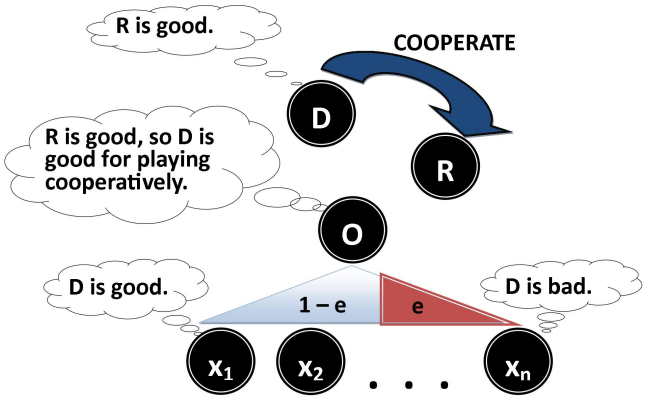


Figure 3-1: During a round of the donation game the donor (D) either cooperates (gives a cost c) with the receiver (R) (gains a benefit b), or defects (no cost paid). The donor's behaviour is judged by an observer (O) based on a norm, and its reputation is altered accordingly. The observer then spreads the newly formed reputation to the rest of the agents ($x_1 ... x_n$). However, the reputation may be spread with some amount of error.

cultural topographies. They found that "folktales from the same culture found 100 km apart are, on average, as similar as folktales found 10 km apart in different cultures" (Ross et al., 2013). By adjusting the levels of value homophily, we consider the advantages of employing accurate personal information over inaccurate social information.

3.4 Model 1: Cooperation via Indirect Reciprocity

3.4.1 Simulation

Here we demonstrate the evolution of cooperation using the donation game. This experiment replicates what is already well-known; reputation can sustain cooperation even against error in communication (Panchanathan and Boyd, 2003). We use this as a baseline for judging the sustainability of cooperation in the subsequent experiments. As such, we have chosen parameters in line with previous instantiations of the donation game so that subsequent experiments are comparable to the existing literature.

In a round of the donation game an individual has the opportunity to pay a cost to give some other agent a benefit. Initially, there are two players, and the *donor* decides to cooperate or defect with the *receiver*. If cooperation is selected, the donor pays a cost c to give a benefit b to the receiver. If the donor defects, the pay-off is zero for both players. As a result, the donating agent may garner a positive (or negative) reputation and be aided (or not) by someone else at some later date.

Symbol	Parameter	Value	Notes
n	population size	100	number of agents.
r	rounds	10,000	donation games per generation.
g	generations	500	# of evolutionary iterations.
b	benefit	[1 - 6]	incremented by 0.5
c	cost	1	cost for cooperating. kept constant.
u	mutation rate	0.01	chance of strategy mutation.
h	homophilly	0, 0.5, 1	see main text.
e	error	0–1	see main text.

Table 3.3: Table of free parameters, values used in present figures, and a sensitivity report including range of values tested.

We use a three-player version of the donation game. Here, an additional player (the *observer*) is permitted to monitor the interaction (see Figure 3-1). The observer then spreads its reputational judgement of the donor. This paper assumes an observer in all interactions, and that the observer judges the reputation of the donor via the 'judging' social norm (though we explore other norms in the appendix).

An observer alters its belief in the donor's reputation via a pervasively accepted social norm, referred in the literature as the Judging norm (Nakamura and Masuda, 2012). When observers employ this norm, we know cooperation can be sustained via indirect reciprocity (Ohtsuki and Iwasa, 2006, 2004; Uchida and Sasaki, 2013; Uchida and Sigmund, 2010). Table 3.2 depicts the Judging norm. If the donor elects to cooperate and the observer believes the recipient's reputation to be good, the observer will assign the donor a good reputation. Conversely, if the donor cooperates, but the observer believes the receiver to be bad, then the observer marks the donor as bad. The reverse is true if the donor defects. The observer then shares its new reputational view of the donor with the rest of the population, and they update their beliefs accordingly (see Figure 3-1).

As per Nowak and Sigmund (1998), three strategies are considered: *always defect* (ALLD), *always cooperate* (ALLC), and *discriminate* (DISC). As donors, discriminating (DISC) agents cooperate or defect based upon their belief in the reputation of the receiver. If the donor believes the agent to be good, it will cooperate, otherwise it will defect. The other two strategies do not consider the recipient's reputation, but always cooperate or defect, as per their namesake.

Error (e) is limited to the range [0...1] and represents the percent of the population which receives the wrong reputational adjustment from the observer. If $e = 0$, then the observer convinces the entire population of the donor's new reputation. Otherwise, some fraction of the populace receives the wrong reputation. For instance, if $e = 0.2$, then 20% of the population inaccurately updates their reputation for that donor (see Figure 3-1). Which individuals receive the erroneous reputation update is selected at random during each interaction. Thus, one agent might possess inaccurate information for Agent A, but not Agent B.

We assume that there is a population of $N = 100$ individuals each with a single triallelic loci representing strategy and a list (of size $N - 1$) representing their *belief* about the reputation of every other individual.

Each element of the list exists in one of two states, which we call *good* and *bad*. In order to test the sustainability of a cooperative society, initially the population is comprised solely of DISC agents. We then analyze whether cooperation can be maintained in the face of invading ALLDs, or indirectly, via ALLCs which are vulnerable to ALLDs.

During each round (r), an instance of the donation game occurs with three randomly selected agents taking on the role of recipient, donor, and observer. Based on its strategy, the donor will choose whether or not to donate to the receiver. If the donor cooperates, it will pay a personal cost c to give a benefit of b to the receiver. Based on the donor's behaviour and the observer's belief of the reputation of the receiver, a fraction of the population $1 - e$, will change their reputational belief of the donor to sync with the observer's new belief. The rest of the population (e), will erroneously update their belief to be the opposite of the observer's.

These interactions are repeated for $r = 10,000$ rounds, constituting a single generation. At the end of a generation, the total pay-off (i.e. the sum of all the costs paid and benefits received of each individual during the generation) is calculated. This represents an individual i's fitness (p_i), which is used to calculate the proportion of the various strategies in the subsequent generation. To do this, we employ the evolutionary selection algorithm, fitness proportionate selection (Goldberg and Deb, 1991). In fitness proportionate selection, each member of the new generation is determined by selecting an individual from the previous generation, with the chance of selecting individual i being $p_i/(\sum_{j=1}^{n} p_j)$, the payoff of individual i divided by the total pay-off of the population. There is then a small chance for mutation $u = 0.01$, where an agent's strategy morphs to one of the three available strategies. The experiment runs for $g = 500$ generations (see Table 3.3 for parameters and sensitivity).

Results

Figure 3-2 illustrates the results, displaying the average fraction of cooperative actions per generation over the final fifty generations. The figure displays the fraction of cooperative acts through a parameter sweep of error rate (e) and benefit/cost (b/c) ratio. Clearly, there exists a threshold of discontinuity. In the white area of the curve the vast majority of actions are cooperative, whilst in the dark area cooperation destabilizes and ALLDs invade. The outlier at $b = 1.5$ and $e = 0$ illustrates that occasionally, with low error rates, ALLCs invade and are subsequently vulnerable to ALLDs (Fishman, 2003; Ohtsuki and Iwasa, 2004). Generally, the figure illustrates that as the benefit/cost ratio increases, a cooperative society can support increased error in reputational dissemination. However, regardless the benefit/cost, cooperation seems to be limited by an error rate of approximately 0.2.

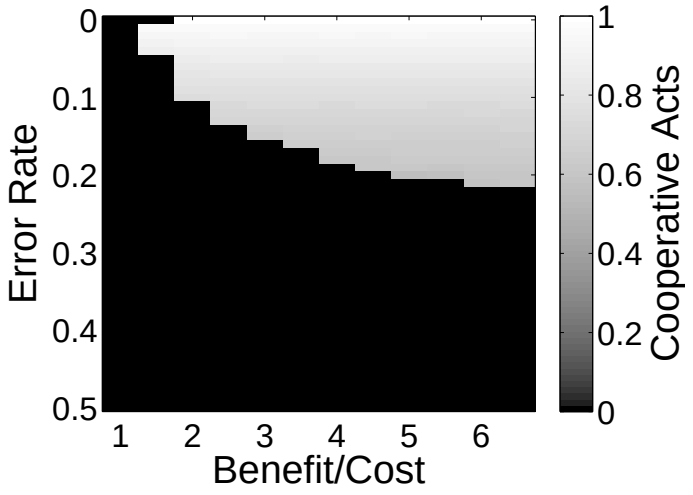


Figure 3-2: Average fraction of donation games which were cooperative over final 50 of 500 generations. A parameter sweep of the error rate and benefit/cost ratio illustrates that the evolutionary sustainability of cooperation depends on the interplay of both parameters. See Table 3.3 for complete details of parameters.

3.5 Model 2: Homophily

3.5.1 Simulation

Model 1 shows that when agents randomly interact, selection for cooperation is limited by error in communication. However, in human societies random interaction is uncommon, a correlation is frequently found between the beliefs of individuals and their propensity to interact (Adamic et al., 2003; Curry and Dunbar, 2013; Fiore and Donath, 2005; Kandel, 1978; McPherson et al., 2001). In Model 2 we use both computer simulations and analytical analysis to examine the effects of this phenomenon, called *value homophily*.

To analyse the effects of value homophily on the evolution of pro-social behaviour, we define it as the correlation between informational relatedness and the propensity to interact. Thus, we seek to juxtapose a randomly interacting society (low value homophily) with a society where individuals tend to interact with like-minded others (high value homophily).

We presume that there is some abstracted cultural variable affecting the distribution of reputation. Namely, an agent will more frequently interact with another agent that shares its reputational assessment of a given recipient compared to a random interaction. Thus, we added a new variable, homophily, $h = [0...1]$.

Rather than defining a specific social structure or network, we operationalised homophily as the probability of a donor interacting with an observer who agrees with the donor's belief in the reputation of the recipient. Initially, for each interaction the donor, recipient, and observer are chosen at random — this remains unchanged. However, if the donor and the observer disagree on the recipient's reputation, then h

represents the probability of replacing the observer with an agent who agrees with the donor's assessment of the recipient's reputation (if such an agent exists). As an example, if $h = 0.2$, then if the donor and observer disagree, there is a 20% chance a different observer will be chosen. This observer will be drawn from the subset of agents that share the donor's belief with respect to the specific recipient.

Importantly, it should be noted that the same amount of error is propagated (and therefore exists) within the population. Error is introduced entirely by observers misinforming e percent of the population. The only difference between Models 1 and 2 is via the selection criterion of the observer. Thus, when the donor is randomly selected, the probability that the donor's belief of the recipient's reputation is erroneous remains unchanged from the first model.

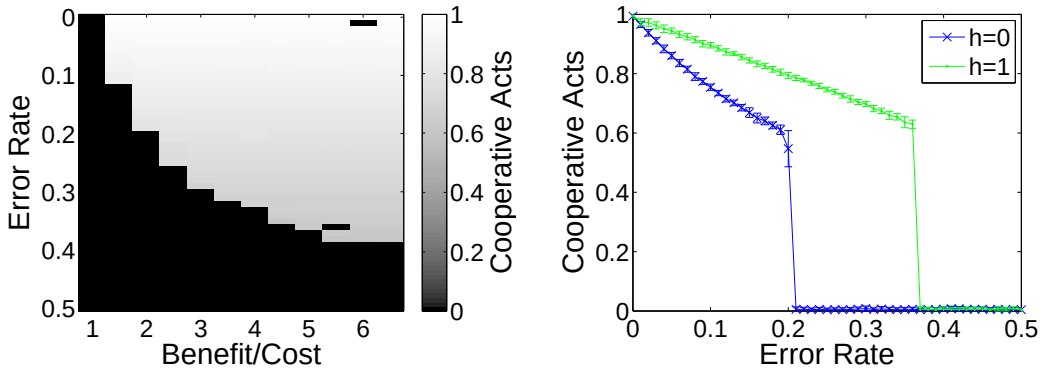
It is also worth noting why we define homophily as congruency between the donor and observer, rather than the donor and recipient. Why not implement value homophily such that a donor is more likely to meet a like-minded recipient? In recent studies of indirect reciprocity involving in-group/out-group dynamics, it is the donor and recipient, rather than the donor and observer, that belong to the same group (Nakamura and Masuda, 2012; Masuda, 2012; Jusup et al., 2014; Matsuo et al., 2014).

We suggest a different way of considering in-group within indirect reciprocity. The observer is as likely to share the donor's social network as the recipient. As indirect reciprocity is founded upon the assumption that agents do not meet twice, we present an example with mobile agents. In Village X, Agent A is well-reputed, but in Village Y, Agent A is ill-reputed. A highly-mobile agent visits Village X, hears the reputation of Agent A, and then meets Agent A. Since the mobile agent is visiting Village X, when it meets Agent A, it has a higher probability of being observed by someone from Village X, rather than from Village Y. Thus, it will likely be judged by Village X's reputational belief. After the interaction, the mobile agent moves on to Village U or W and interacts with Agent B. In this sense, an agent does not have a group with which it interacts, but when it interacts with an agent, it is observed by the agents who informed the wandering agent of the recipient's reputation. We will return to this point in the discussion and test the consequences of defining homophily differently.

Results

Figure 3-3 illustrates the consequences of homophily on the sustainability of cooperation. If one juxtaposes Figure 3-3(a) — which shows the maximum degree of homophily ($h = 1$) — with Figure 3-2, where $h = 0$, it is clear that given unanimity between the donor and the observer, stable cooperation can tolerate greater amounts of error. In Figure 3-2 cooperation fails at an error rate of approximately $e = 0.2$. With homophily, Figure 3-3(a), cooperation can sustain an error rate of around $e = 0.4$. Furthermore, this result is reproducible in a variety of environments. In A.1.1 we show that this result extends to other social norms (e.g. Standing and Shunning). In A.2 the result is duplicated when reputation is not binary, but more continuous.

Figure 3-3(b) depicts a subsection of the results underlying Figures 3-2 and 3-3(a), where the benefit/cost



(a) Average fraction of cooperative acts.

(b) Comparing $h = 1$ with $h = 0$.

Figure 3-3: **(a)** Depicts the average fraction of donation games which were cooperative over the final 50 generations. Homophily is 100% ($h = 1$). **(b)** Comparison of the fraction of cooperative acts with $h = 1$ and $h = 0$. Benefit/Cost is held static at 5. Error rate varies [0,0.5] in increments of 0.01

ratio is 5. The graph shows the average fraction of cooperative acts for both full homophily ($h = 1$) and the original instantiation of randomly selected observers ($h = 0$). For a given error rate, the number of cooperative acts is increased when the donors and observers agree upon the recipient's reputation, regardless of the accuracy of the reputation. When $h = 0$ cooperation fails at $e \approx 0.2$, while unanimity in reputational assessment allows stable cooperation to survive twice the error rate ($e \approx 0.4$). Furthermore, not only does homophily sustain cooperation against more error, when cooperation is sustained for either society, a homophilous society performs more cooperative acts. For example, at $e = 0.1$ cooperation is stable for both homophilous and non-homophilous societies, however more cooperative acts take place when $h = 1$.

3.5.2 Simplified Analytical Model

With simulations it is difficult to elucidate the mechanisms underlying the result. Here we attempt to analytically model the repercussions of homophily on cooperation. We make a few simplifying assumptions from the simulation, though they should not seriously impact the result. First, we presume that an agent will first be selected as a donor and then as a recipient. Second, we only consider two strategies, DISC and ALLD. With the Judging norm and error in communication, there is a relatively low danger of ALLC agents drifting into the population sufficiently to risk a subsequent ALLD invasion (Takahashi and Mashima, 2006; Fishman, 2003; Ohtsuki and Iwasa, 2004). Consequentially, we focus on when ALLDs can invade a society of DISCs, and how a homophilic society guards against it. Finally, we do not employ evolution, rather we only test when rare ALLDs perform better than a population comprised almost entirely from DISCs.

We presume an infinite population where donation games are played for a finite, but extended period of time. We define P_r and P_d as the probabilities of interacting with a DISC (i.e. reciprocater) and an ALLD,

respectively. After each game, w is the probability that another round of the donation game will take place. Similar to Panchanathan (2011), we presume that the proportion of reputations goes to equilibrium by the second round. Given this, the fitness of an ALLD (π_d) may be expressed as:

$$\pi_d = bP_r + \frac{w}{1-w}(bP_rG_d) \tag{3.1}$$

In the first round, each agent is considered good, so an ALLD agent will receive a benefit if it meets a DISC agent (bP_r). In subsequent rounds, the ALLD agent will receive a benefit if it meets a DISC agent, and the DISC agent believes the ALLD agent is good. G_d represents the probability that another agent considers the ALLD agent good.

The fitness of a DISC agent may be expessed by equation 3.2. In the first round, the agent will receive a benefit if it interacts with another DISC. Additionally, it will always pay the cost of cooperating as each agent begins with a good reputation ($bP_r - c$). In subsequent rounds, it will receive a benefit if the agent interacts with another DISC, and that agent believes the DISC to be good (G_r). It will pay the cost of cooperating if it interacts with any agent who it believes to be good.

$$\pi_r = bP_r - c + \frac{w}{1-w}[bP_rG_r - c(P_rG_r + P_dG_d)] \tag{3.2}$$

In our simulation we test whether a population of DISCs is stalwart against invasion. In the analytical model, this would be true as long as the average fitness of DISC agents is greater than the average fitness of ALLD agents:

$$\pi_r - \pi_d > 0 \tag{3.3}$$

Furthermore, in the invasion scenario, we presume that initially $P_r \approx 1$ and $P_d \approx 0$. Presuming play continues for an extended period ($w \approx 1$), we expect a population of DISCs to be stable against invasion if (see Appendix A.4.1 for details):

$$G_r > \frac{bG_d}{b-c} \tag{3.4}$$

Equation 3.4 shows that the stability of a DISC population is dependant on the likelihood that agents of a strategy are well-reputed. The probability of having a good reputation not only depends on a donor's actions, but on the error rate of reputational dissemination, and how likely agents with different reputational beliefs interact. This is how homophily alters the interactions of a population.

In a randomly interacting population ($h = 0$), the probability of a DISC agent possessing a good reputation is:

$$G_r^{h=0} = (1-e)[(1-e)^2 + e^2] + e[2e(1-e)] \tag{3.5}$$

Given the Judging social norm, a donor DISC agent earns a good reputation if the observer and it either: *a*) agree on the reputation of the recipient, and the reputation is passed on correctly, or *b*) disagree on the reputation of the recipient, but the reputation is passed on erroneously.

In the first case, there are two ways the donor and observer can agree on the reputation of the recipient. Either, both possess correct information about the recipient $(1-e)^2$, or they both possess incorrect social information, e^2. The good reputation is then disseminated to $(1-e)$ fraction of the population. In the second case, the donor and observer disagree with probability $[2e(1-e)]$, and the DISC agent receives a bad reputation. However, e fraction of the population erroneously believe the donor is good.

In a population with homophilic interactions (h), the probability of a DISC agent possessing a good reputation is altered. Equation 3.5 becomes:

$$G_r = h(1-e) + (1-h)G_r^{h=0} \tag{3.6}$$

If $h = 0$ then Equation 3.6 = Equation 3.5, since each donation game is played with randomly selected agents. In a fully homophilous society $h = 1$, the donor and the observer will always agree, therefore the DISC agent will always earn a good reputation. However, because there is error in transmission, only $1-e$ fraction of the population will believe the donor is good. Thus, $G_r = 1-e$.

The probability of a ALLD agent receiving a good reputation (G_d) is:

$$\begin{aligned} G_d = (1-e)[P_r(1-G_r) + P_d(1-G_d)] + e[P_r(G_r) + P_d(G_d)] \\ G_d = (1-e)(1-G_r) + eG_r \end{aligned} \tag{3.7}$$

An ALLD always defects, so it will receive a good reputation if the observer believes the recipient is bad. If the recipient is a DISC agent, then the observer will believe the agent is bad with probability $(1-G_r)$. Similarly, if the recipient is an ALLD, then the observer will hold it negative repute with probability $(1-G_d)$. The observer will then spread this good reputation to $(1-e)$ of the population. Furthermore, if the observer believes the agent is good, it will give the ALLD donor a bad reputation, but e fraction of the population will erroneously consider the agent good. However, since the assumption for invasion is that $P_r \approx 1$ and $P_d \approx 0$, the equation simplifies to $G_d = (1-e)(1-G_r) + eG_r$.

The probability an agent will consider an ALLD agent good (G_d) does not change based on homophily. In a homophilous interaction a donor and recipient are chosen at random. The observer is then selected to agree with the donor, but the donor's opinion of the recipient is selected from the same probability as the observer's opinion of the recipient. However, while Equation 3.7 remains the same for any value of homophily, the calculation of G_r changes, so in that sense, G_d is altered.

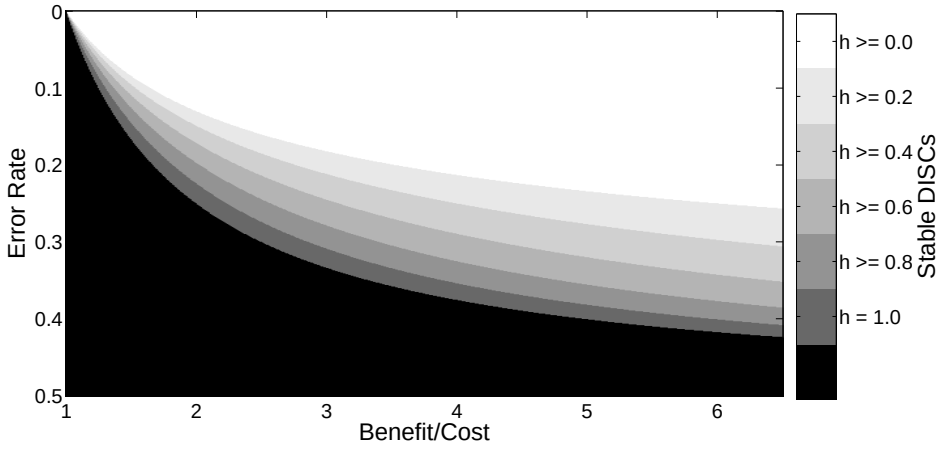


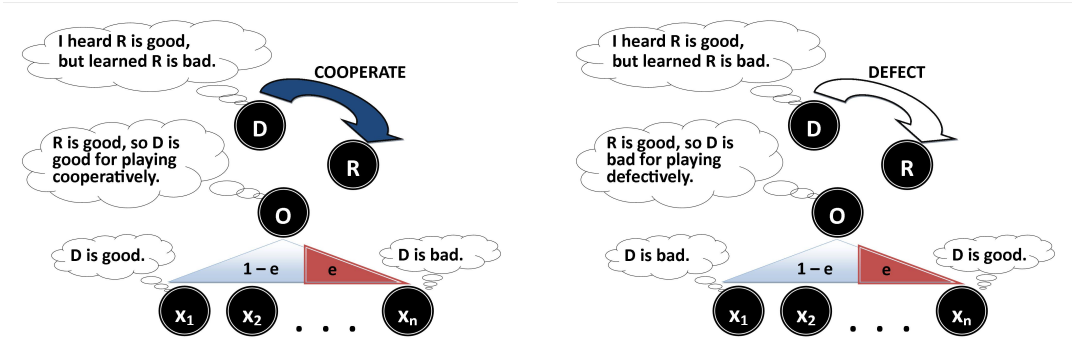
Figure 3-4: Plotting Equation 3.4 for differing values of homophily. The **black** area shows where ALLDs are predicted to invade. The **grey** areas represent where, for certain values of homophily, DISCs are stable against invasion. Lighter greys are subsets, of darker greys. For example, the area of DISC stability for $h = 0.0$ is a subset of $h = 0.2$. As the frequency of homophilous interactions increase, DISCs can avoid invasion despite greater frequency of errors in communication.

Results

Plotting $G_r > \frac{bG_d}{b-c}$, Figure 3-4 illustrates where the model predicts DISC stability. A parameter sweep of benefits (b) and error rates (e) are tested. The black area defines where $G_r \leq \frac{bG_d}{b-c}$, and thus, where we predict an ALLD invasion. The grey areas depict parameter locations where $G_r > \frac{bG_d}{b-c}$ for certain values of h, and thus, where a DISC population is stable.

Figure 3-4 is in sync with the simulation results. The whitest curve in Figure 3-4 is comparable to Figure 3-2, and the darkest grey (that is not black) is comparable to Figure 3-3(a). Despite the inclusion of ALLCs and the lack of evolution, the analytic model predicts similar DISC stability.

Of note, with the simulations we showed that cooperation is more robust to error in homophilous societies. Here we demonstrate the reason why. Homophily increases the ratio of good DISCs over good ALLDs (i.e. G_r/G_d). When the benefit (b), cost (c), and error rate (e) are held constant, larger values of h increase the robustness of a cooperative society. In general, both the computer simulations and analytical analysis demonstrate that homophilous societies help sustain cooperation against error in reputational dissemination.



(a) DISC

(b) VDISC

Figure 3-5: Difference between veridical (VDISC) and socially-biased (DISC) agent. The donor (D) has been socially informed that the recipient (R) is good, but has discovered that this information is inaccurate. **(a)** The DISC donor decides to act on societal information. **(b)** The VDISC donor decides to act on accurate information, despite negative normative consequences.

3.6 Model 3: Incorrect Social Information v. Correct Personal Information

3.6.1 Simulation

Lastly, we introduce the conflict between descriptive and normative information. Thus far, we have assumed that reputational errors were spread randomly, without intention. However, while reputation is spread, an individual may realize that its socially-acquired reputational (normative) belief concerning another agent is in error. In such circumstances, is it in the agent's best interest to update its reputational belief and act on its accurate (descriptive) information? Or, is it better off continuing to act upon the inaccurate, socially-informed information?

To model this we added a new binary variable, *veridicality*, v. If $v = 0$, then a donating agent will always employ its socially received reputational belief when selecting whether to cooperate or defect against a recipient. In Models 1 and 2, v was implicitly zero. Donors always acted based on their belief of the recipient's reputation. If $v = 1$, then the donating agent will access and employ correct, descriptive information when interacting with a recipient. To do this, we give the donor access to the reputation the recipient would have garnered if there was no error rate. The recipient can then be judged on its merits rather than on error-prone gossip. While probably not realistic (though see Fetchenhauer and Dunning (2010) and Sigmund (2009)), we test this as the extreme limit case, to understand the impact of having such information freely available. If it is not in an agent's best interest to employ correct information even when it is free, then it would not pay a cost to attain the information.

We now add a fourth agent strategy, *veridical DISC (VDISC)*. VDISC agents behave like DISC agents,

except that when a VDISC agent is a donor, and the agent's reputational belief of the receiver is in error, then the VDISC behaves based on its descriptive knowledge rather than its socially received normative knowledge.

Figure 3-5 illustrates the difference between VDISC and DISC strategies. In both Figure 3-5(a) and 3-5(b), the donor has been socially informed that the recipient is good. Furthermore, the donor has learned, or is able to observe, that this socially-acquired information is in error. Figure 3-5(a) exemplifies the discriminating (DISC) agent. Despite this new information, the agent continues to employ the social information. In Figure 3-5(b), the veridical discriminating agent (VDISC) communicates the accurate information.

To analyse the effects of veridicality, we test whether VDISC agents can be invaded by DISCs. As such, the initial population is comprised entirely of VDISC agents. Mutation now offers the possibility of four strategies entering the population, $S_i \in \{ALLD, ALLC, DISC, VDISC\}$. We observed the consequences across three values of homophily, $h \in \{0, 0.5, 1\}$.

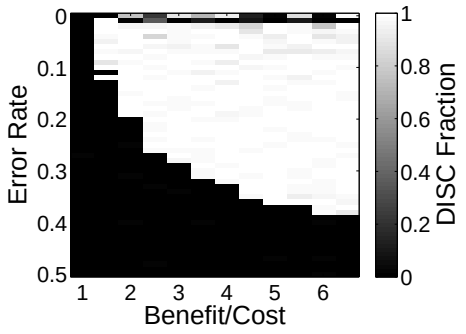
In conditions where $h > 0$, finding a homophilous observer precedes the donor's decision regarding whether to shift its reputational belief to the correct information. Once a donor is selected, if the observer disagrees with the donor's reputational belief of the recipient, a new observer might be selected which agrees with the donor (see the explanation of homophily under Experiment 2). This permits us to pose the question: if one typically interacts with agents who agree but may prove to be in error, is there utility in switching to acting on a known truth?

Results

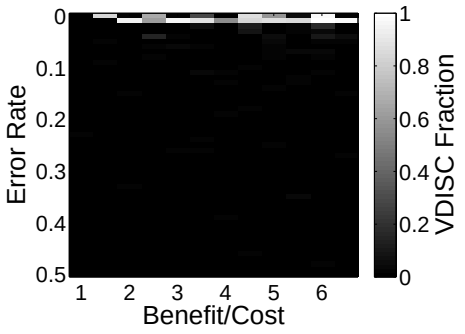
Figure 3-6 depicts the fraction of the population consisting of DISCs and VDISCs after 500 generations for varying values of homophily. Since at generation 0 the entire population was comprised of VDISCs, the figure illustrates that DISCs invade when $h = 1$ or $h = 0.5$ (this is reproduced for the Standing and Shunning norm, but not Image Scoring in A.1.2). Figures 3-6(a) and 3-6(b) exemplify the consequences given full homophily ($h = 1$). In general, agents acting on correct information are invaded by the those acting upon socially garnered information. When $h = 0.5$, the DISC strategy still invades the VDISC agents, but to a lesser extent (Figures 3-6(c) and 3-6(d)). Finally, Figures 3-6(e) and 3-6(f) show that in an environment where the interactions are completely independent from the beliefs of a given recipient ($h = 0$), VDISCs are stable against DISCs.

3.6.2 VDISC Stability: Simplified Analytical Model

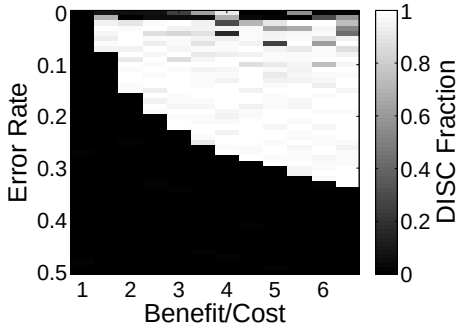
Next, we employ an analytical approach to understanding when DISCs can invade a population of VDISCs. We do not consider ALLCs and ALLDs, focusing on the interaction between the two reputation considering strategies. Most of the assumptions remain unchanged from Section 3.5.2. $w \approx 1$ is the chance of playing another round. Each agent starts with a good reputation, and we presume the percent of well-reputed agents equilibrate by round 2. The population begins with mostly VDISCs and few DISCs, such that the proba-



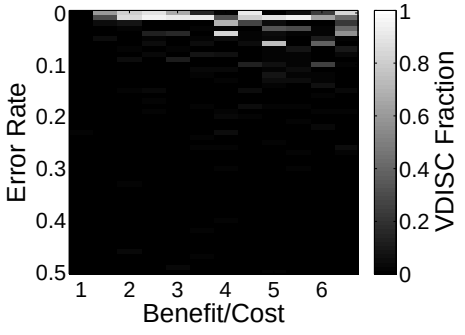
(a) Fraction of DISC agents when $h = 1$.



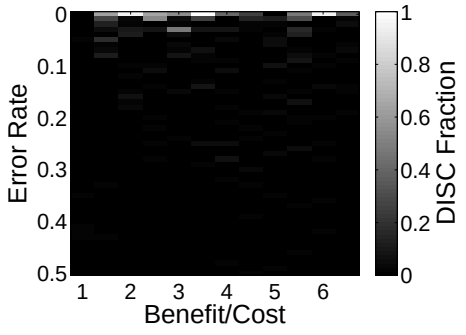
(b) Fraction of VDISC agents when $h = 1$.



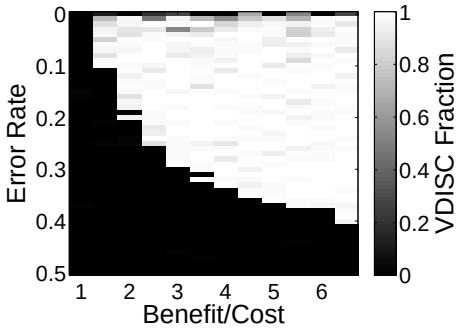
(c) Fraction of DISC agents when $h = 0.5$.



(d) Fraction of VDISC agents when $h = 0.5$.



(e) Fraction of DISC agents when $h = 0$.



(f) Fraction of VDISC agents when $h = 0$.

Figure 3-6: Fraction of population composed of a strategy after 500 generations. The initial population is solely comprised of VDISCS.

bility of interacting with a VDISC is approximately one ($P_v \approx 1$), and the chance of meeting a DISC is approximately 0 ($P_r \approx 0$).

A DISC agent's fitness (π_r) can be defined as:

$$\pi_r = b - c + \frac{w}{1-w}[b(P_r G_r + P_v G_r) - c(P_r G_r + P_v G_v)] \tag{3.8}$$

In the first round each agent is positively reputed, so a DISC both receives a benefit and cooperates ($b - c$). In subsequent rounds, a DISC agent receives a benefit if it interacts with a DISC or VDISC who considers it good, $b(P_r G_r + P_v G_r)$. G_v and G_r are the chance an agent will consider a VDISC and DISC to be good, respectively. Furthermore, the DISC agent pays a cost if it interacts with any agent it considers good $c(P_r G_r + P_v G_v)$.

A VDISC agent's fitness (π_v) is:

$$\pi_v = b - c + \frac{w}{1-w}[b(P_r G_v + bP_v G_v) - c(P_r G_r + P_v G_v)] \tag{3.9}$$

Similar to the DISC, in the first round a VDISC will both receive a benefit and cooperate ($b - c$). In subsequent rounds, it will receive a benefit if it meets an agent who considers it good. It will pay a cost and cooperate if it meets an agent it believes to be good. To test whether VDISCs can repel an invasion of DISCs, we analyze when the fitness of a VDISC agent is greater than that of the DISC:

$$\pi_v - \pi_r > 0 \tag{3.10}$$

Since we presume $P_v \approx 1$, $P_r \approx 0$, and $w \approx 1$, Equation 3.10 simplifies to (see A.4.2 for details):

$$G_v > G_r \tag{3.11}$$

If the chance that a VDISC is considered good is greater than that of a DISC, then the population of VDISC is stable. Next, we show that both G_v and G_r are functions of error rate (e) and homophily (h). We then analyze the interplay between both on the stability of a VDISC population.

In a society with random interactions ($h = 0$), the fraction of good VDISC agents is:

$$G_v^{h=0} = [P_v + P_r(1-e)](1-e) + (P_r(e))^2 \tag{3.12}$$

A VDISC, like a DISC agent, receives a good reputation if it agrees with the observer on the reputation of the donor. Since the VDISC donor always switches to the correct information, it will agree with the observer with probability $1 - e$. Furthermore, when the VDISC becomes a recipient, this good reputation is accurately used by $[P_v + P_r(1-e)]$. Only $(1-e)$ fraction of the population hears the correct social information, however, all VDISCs will employ the correct reputation. Lastly, if the observer is in error, then the VDISC and observer will disagree, garnering the VDISC a bad reputation. However, when the VDISC

becomes a recipient, it will receive a benefit if the donor is a DISC who is in error of the original error. Since we presume $P_v \approx 1$ and $P_r \approx 0$, the equation simplifies to:

$$G_v^{h=0} = 1 - e \tag{3.13}$$

Adding homophily, the chance that a donor agent who interacts with a VDISC recipient will consider the recipient good, is:

$$G_v = h(1-e)[P_v + P_r(1-e)] + (1-h)G_v^{h=0} \tag{3.14}$$

In a homophilous interaction a VDISC donor will be matched with an observer that heard the same social information regarding the reputation of the recipient. Either, both the donor and recipient heard the correct social information with probability $1 - e$, or both heard incorrect information with probability e. However, the VDISC donor then discovers the correct reputation of the recipient, and employs that. Thus, it will only agree with the recipient with probability $1 - e$. This reputation will then be employed by a donor with probability $[P_v + P_r(1 - e)]$. The correct, good reputation will always be employed by a VDISC, and will be employed by a DISC if it heard the correct information through gossip, $P_r(1 - e)$. Presuming $P_v \approx 1$ and $P_r \approx 0$, the equation simplifies to:

$$G_v = h(1-e) + (1-h)(1-e) \tag{3.15}$$

For a DISC agent in a randomly interacting population ($h = 0$), the probability of possessing a good reputation is:

$$G_r^{h=0} = [P_v + P_r(1-e)][(1-e)^2 + e^2] + P_r e[2e(1-e)] \tag{3.16}$$

A donating DISC will be treated as if it has a good reputation if the observer agrees with it regarding the reputation of the recipient $[(1 - e)^2 + e^2]$, and then, as a recipient, the agent interacts with a VDISC donor (who will act on the correct information), or a DISC donor who heard correct gossip ($[P_v + P_r(1 - e)]$). Additionally, it will be considered good if it interacts with an observer who disagrees $[2e(1 - e)]$, but as a recipient meets a DISC donor in error $(P_r e)$. In the invasion case, this simplifies to:

$$G_r^{h=0} = (1-e)^2 + e^2 \tag{3.17}$$

Adding homophily:

$$G_r = h[P_v + P_r(1-e)] + (1-h)G_r^{h=0} \tag{3.18}$$

If $h = 0$ then Equation 3.18 = Equation 3.16, since each donation game is played with randomly selected agents. In a fully homophilous society $h = 1$, the donor and the observer will always agree, therefore the DISC agent will always earn a good reputation. However, because there is error in transmission, only

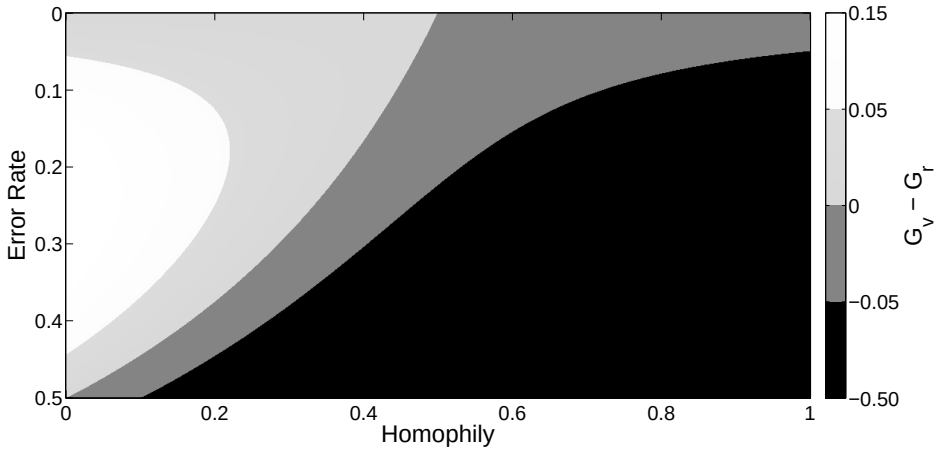


Figure 3-7: Plot of $G_v - G_r$ over a parameter sweep of error rate (e) and homophily (h). A positive difference represents the phase space where a population of VDISCs is stable against invasion from DISCs. A negative difference suggests invasion. **White** represents strong selection for VDISCs. **Light Grey** is weaker selection for VDISCs. **Black** represents strong selection for DISCs. **Dark Grey** illustrates weaker selection for DISCs. The delineation between light and dark grey marks where VDISC and DISC fitness are equal.

$P_v + P_r(1-e)$ fraction of the population will believe the donor is good. Again, presuming $P_v \approx 1$ and $P_r \approx 0$, and substituting in $G_r^{h=0}$, the fraction of good DISC agents is:

$$G_r = h + (1-h)[(1-e)^2 + e^2] \tag{3.19}$$

According to Equation 3.11, a VDISC population is stable against a DISC invasion if Equation 3.15 - Equation 3.19 > 0. Substituting, we get:

$$h(1-e) + (1-h)(1-e) - (h + (1-h)[(1-e)^2 + e^2]) > 0 \tag{3.20}$$

Results

Graphing the equation, Figure 3-7 depicts the phase space where VDISCs are stable against DISCs. The Figure represents Equation 3.11, reduced to a function of e and h, by Equation 3.20. It is plotted over a parameter sweep of error rate (e) and homophily (h). A positive difference represents where a population of VDISCs is stable against invasion from DISCs. A negative difference suggests invasion.

The curve between the light grey and dark grey region represents where both strategies perform equally. A larger difference, suggests stronger selection, so the graph differentiates where $|G_v - G_r| < 0.05$.

The results explain the simulation result in Figure 3-6, where VDISCs are stable against invasion without

homophilic interactions ($h = 0$), but are invaded when $h = 0.5$ or $h = 1$. As homophily increases, VDISCs require higher fidelity communication in order to repel invasion. However, when $h \geq 0.5$, any error in communication favors DISC agents.

It should be noted that our simple analytical model only compares the fitness of VDISCs to DISCs when the population is almost entirely comprised of VDISCs. The simulation offers a more complicated, evolutionary solution. It takes into account the repercussions of DISCs increasing in frequency, as well as interacting with ALLD and ALLC agents. This is why in the simulation results (Figure 3-6) both VDISCs and DISCs are still invaded by ALLDs agents for certain error rates.

3.7 Discussion

We have presented three experiments, the first two demonstrated results that are becoming increasingly well-known. First, cooperation is adaptive provided that the overall benefits outweigh the costs. Second, many kinds of social structures, including value homophily, extend the range of cost/benefit ratios over which cooperation can flourish or even fixate. Unjustified reputations can threaten cooperation, and previous studies have sought a method for removing false reputations from a population (Nakamaru and Kawata, 2004). The present paper tested the percentage of false reputations which threatens cooperation, and found that stratifying the interaction propensity of those with correct and incorrect information, augments cooperation.

The third experiment though shows a truly novel result. As value homophily increases, not only does the robustness of a cooperative society increase, but, paradoxically, so does selective pressure for employing error-prone social information — even when correct information is freely available. As normative pressures rise, cooperation is enhanced, but there is a consequent pressure for conformity, meaning that the relative utility of descriptive information decreases. Where there is no value homophily, descriptive (veridical) information about the strategies of others holds the advantage. But where value homophily is deployed, conformity to norms is more adaptive than acting on the truth, at least in the context of these simple agents.

3.7.1 Gossip

Gossip is widespread in humans (Dessalles, 2007) and yet is also widely disparaged. Gossips were burnt in medieval Europe (Emler, 1990), and women, in particular, have been repressed to avoid it (Funder, 1995; Gilmore, 1978). Intuitively, this is understandable: gossip is often false, conspiratorial, unverifiable and malicious. Yet evidence suggests that gossip is pervasive in its influence on human behaviour (Ellwardt et al., 2012). Further, empirical research suggests that much gossip focusses on the true and positive (Ellwardt et al., 2012; Herrmann et al., 2012; Ingram and Bering, 2010; Sommerfeld et al., 2008), and theoretical results, including those reported here, show that it can be beneficial for spreading information that allows individuals to choose who to interact with (Giardini and Conte, 2012; Mitchell et al., 2013; Nakamura and Masuda, 2011; Traag et al., 2011). Our results suggest one way in which the seeming contradiction between

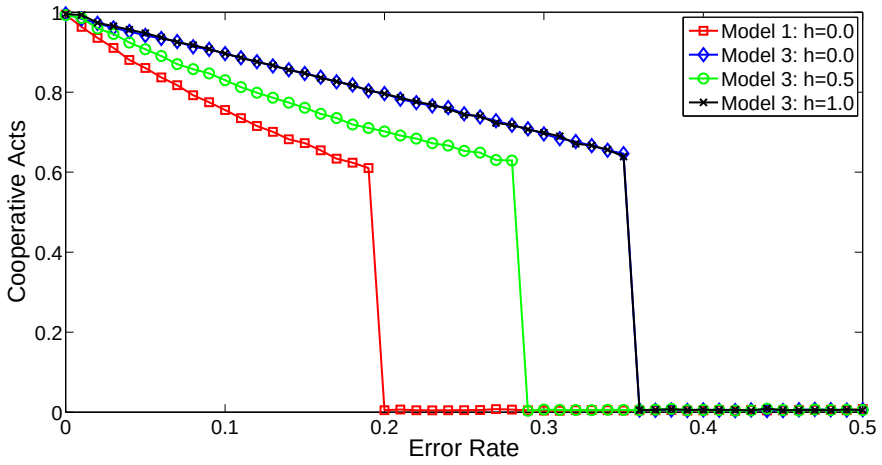


Figure 3-8: Comparing the consequences of Experiment 3 with cooperation. Each line represents the fraction of cooperative actions when benefit/cost was held at 4.5. In homophilous societies, agents who employ social information in lieu of veridical information sustain cooperation similarly to veridical agents who interact without any abstracted cultural bias ($h = 0$).

gossip's negative capacities and its pervasive presence might be resolved. Behaviour which is occasionally unfair to individuals who are the object of incorrect gossip delivers an overall benefit.

Giardini and Conte (2012) posit that while gossip does not necessarily transmit accurate descriptive information, it likely imparts what others believe is the descriptive information — deemed meta-evaluation. In a human experiment of the donation game, Sommerfeld et al. (2008) demonstrated that gossip (i.e. the recipient's reputation) is a better predictor of a donor's decision to cooperate or defect than descriptive information (i.e. the history of the recipient's actions). They posit that individuals might "adjust their own behaviour in order to not depart from the public opinion of their local group; they do not want to stand out" (Sommerfeld et al., 2008). Our model offers an explanation of such a behaviour, namely that meta-evaluation *is* descriptive information and should be considered since the normative pressures embedded in gossip may outweigh the utility of reputational descriptive information.

3.7.2 Cooperation

Whilst we showed that DISCs invade VDISCs at high levels of homophily, we did not discuss the effects of the two strategies on cooperation. Figure 3-8 depicts the fraction of cooperative actions for each condition in Experiment 3. When $h = 0$, VDISC agents sustain cooperation against error just as effectively as DISC agents do when $h = 1$. This is despite the fact that in both conditions, VDISCs exploit descriptive (veridical) information, while DISCs use error-prone gossip.

It could be suggested that rather than generating robustness of cooperation via homophily (which breeds

agents that ignore correct information), society could breed randomly interacting agents that employ correct information. Whilst this is true, we posit that in the natural world correctness of information typically comes at a high cost. Even though, for our model, we offered accurate information without cost, this was to illustrate that DISCs invaded even in such an unrealistically extreme context. We do not believe cost-free descriptive information is feasible in the natural world. In contrast, there is significant real world evidence for value homophily (Adamic et al., 2003; Curry and Dunbar, 2013; Fiore and Donath, 2005; Kandel, 1978; McPherson et al., 2001). Our results contribute to mounting evidence for the utility of homophily in generating cooperative behaviour, and more generally for exploiting and developing novel adaptive traits (Dieckmann and Doebeli, 1999; Panhuis et al., 2001; Powers et al., 2011).

Furthermore, while personally gathered information may be less noisy than socially garnered knowledge, it is certainly not free from error. In our model, personal information was always veridical, and thus agents employing personal information were given an infeasible benefit. Again, this was employed as the limiting case. If this constraint was lifted and error in personally gathered information was included in the model, this might reduce the fitness of the descriptively-biased agents. In terms of cooperation, we have shown two strategies that are equally effective in theory, but only one of which is plausible in practice.

3.7.3 In-Group and Indirect Reciprocity

Humans tend to treat individuals within and without of their group distinctly. Recently, indirect reciprocity research has begun to study in-group biases (Nakamura and Masuda, 2012; Masuda, 2012; Jusup et al., 2014; Matsuo et al., 2014). There are two major differences between our model and the current literature. First, the existing models presume predefined group structure. Second, as mentioned earlier, in-group dynamics are defined by the likelihood that donors and recipients are within the same group, rather than donors and observers. We discuss the consequences of both, in turn.

Nakamura and Masuda (2012) presume group structure, where each group of agents agree upon the reputations of other agents. They show that this and the Judging norm are sufficient for the development of strong *in-group favoritism*. Similar to our model, individuals in a group always agree on the reputation of agents, though they may be in error. So, when group members interact (even if the agreed upon reputation is in error), DISC agents receive a good reputation. This leads to in-group members possessing a higher fraction of good reputations between each other, relative to out-group agents, deemed in-group favoritism.

The present model does not presume group structure. In a homophilous society, if an agent believes Agent A is good, it will interact with others who believe Agent A is good. If it then meets a bad recipient, Agent B, it will be observed by a completely different subset of the population who believes Agent B is bad. In our model there is not a subset of agents with which one agent shares all information. This can be thought of like the analogy of the highly-mobile agent presented in Section 3.5. An agent does not have a group with which it interacts, but when it interacts with an agent, it is observed by the agents who informed the wandering agent of the recipient's reputation.

The second difference between this model and the rest of the literature, is that in-group behaviour is defined by the relationship between the donor and recipient. In our model, the recipient is always chosen randomly, and it is the donor and observer who agree. In A.3 we analyze the repercussion to a society if homophily is defined as unanimity between *a*) the donor and recipient; or, *b*) the observer and recipient. In both cases, the redefined homophily neither benefits or hinders cooperation. Given these differences, we believe it would be interesting to see the consequences to Nakamura and Masuda (2012)'s model if an individual learned the group reputation was wrong and attempted to switch to the correct reputation.

3.7.4 Self-Deception

A society where cooperation is aided by not communicating the truth seems a dire world. But, do we actually see this in the natural world? Do individuals knowingly communicate falsity to align with the group? Our model is agnostic on the intentionality of the agent. What matters is the behaviour (i.e. aligning with the normative information), not whether the agent intends to deceive. As such, this research may have implications for the research on self-deception. von Hippel and Trivers (2011a) argue that intentional deception is cognitively demanding, and that self-deception might evolve to mitigate that cost. Two naturally observed mechanisms which could enable self-deception are (i.) confirmation bias and (ii.) groupthink.

Confirmation bias is a phenomenon whereby individuals bias their information gathering to retain the fidelity of their existing beliefs (Nickerson, 1998; Jonas et al., 2001; Schulz-Hardt et al., 2000). If, as in our model, it is not beneficial to employ correct information, an agent may attempt not to learn the truth, and thus never be confronted with the conflict which possessing correct information generates. Second, if it is easier to deceive another if one does not know one is deceiving the other (Trivers, 1991), then it may be better for an agent not to have explicit access to some knowledge — that is, it should not be able to act on that knowledge, even if it is still acquiring it 'unconsciously' (Bryson, 2009). This seems associated to the phenomenon of groupthink, where being in a group biases the way individuals process information (Janis, 1971; Turner and Pratkanis, 1998). In group contexts, individuals might not be aware of the information they have or the opinions they would have developed if uninfluenced by the group (Janis, 1972; Bénabou, 2013). While deceiving others can be advantageous, the cost of detection is likely high. If accurate information performs worse than inaccurate, and if deception comes at a cost, then equilibrium may rest at a point where the agent would be willing to pay some cost to not acquire the information in the first place (i.e. confirmation bias), or to not be able to access it in some behavioural contexts (i.e. groupthink).

3.7.5 Parsimonious Causes for Ignoring Correct Information

Finally, the correlation between this simple model and the natural world is clearly speculative. However, the donation game has been used as an existence proof for the evolution of cooperation via the spread of reputation (Nowak and Sigmund, 1998). We show that taking this simple model and adding one additional constraint (i.e. value homophily) leads to a phenomenon seen in the natural world. More variance on

whether an individual will donate to another is explained by using error-prone reputation rather than an accurate record of a recipient's previous actions (Sommerfeld et al., 2008). At the very least, we believe this model increases insight into the variables required to give rise to a phenomenon witnessed in the natural world.

3.8 Conclusion

We make two claims regarding the role of homophily and misinformation in employing the donation game to sustain societal level cooperation. First, increasing homophily facilitates cooperation. Even when information is communicated with a high probability of error, social information is better able to sustain cooperation. Second, if society employs homophily, then when faced with a choice between accuracy or group unanimity it is in an individual's interest to prioritize the latter. Therefore, our overall conclusion is that when social norms achieve cooperation there are contexts (e.g. homophilous societies) where persisting with inaccurate normatively-acquired information despite knowledge of its descriptive falsity is advantageous and will tend to evolve.

4

The Evolution of the Impact Bias: Optimizing Affective Forecasts for Decision-Making in Noisy Environments

> *Let us weigh the gain and the loss in wagering that God is... If you gain, you gain all; if you lose, you lose nothing. Wager, then, without hesitation that He is.*

Blaise Pascal, *Pensées*

> *When one is in love, one always begins by deceiving one's self, and one always ends by deceiving others.*

Oscar Wilde, *The Picture of Dorian Gray*

4.1 Summary

Prior to an event, humans systematically fail to predict how they will feel after that event. They predict more intense emotions compared to their actual affective experiences, a phenomenon known as the impact bias. Furthermore, individuals often do not learn to improve their affective forecasts after experiencing this failure.

Even when individuals are given the opportunity to obtain accurate social advice to aid in predicting, they frequently prefer to rely on their own incorrect predictions. Here, we provide a possible explanation. We show that the impact bias is adaptive for optimizing decisions in noisy environments. This result extends to not learning from past experience and social advice. We finish by showing that our parsimonious explanation accounts for significant empirical data.

4.2 Introduction

There is robust evidence that people systematically fail to accurately predict how they will feel about an anticipated experience. Prior to an event, they predict they will be happier for positive events and more unhappy for negative events compared to their actual affective experience (Wilson and Gilbert, 2013; Wilson et al., 2000; Schkade and Kahneman, 1998; Wilson et al., 2003a; Wilson and Gilbert, 2003; Gilbert et al., 1998). This has been referred to as the *impact bias* (Gilbert et al., 2002). Individuals have been shown to predict higher highs and lower lows compared to their eventual experience in the areas of relationships (Eastwick et al., 2008), sporting results (Wilson et al., 2000), life satisfaction (Schkade and Kahneman, 1998), and politics (Gilbert et al., 1998). Not only do individuals fail to predict their affective responses, they do not appear to learn from their mispredictions (Meyvis et al., 2010; Lacey et al., 2006; Scheibe et al., 2011; Wilson et al., 2003b). Furthermore, individuals do not seek social advice to minimize their predictive errors, often employing their own inaccurate predictions even when offered accurate social advice (Gilbert et al., 2009). Though, at first glance, incorrectly predicting affective experiences appears detrimental, given the ubiquity of the impact bias, a recent call has been put forth to discover whether the bias serves some function (Miloyan and Suddendorf, 2015).

Why do we consistently and exaggeratively bias predictions of our emotional experiences? Recent research has demonstrated that the impact bias increases motivation (Morewedge and Buechel, 2013). But, acting on inaccurate affective predictions should lead to suboptimal decisions (Gilbert and Wilson, 2009; Loewenstein, 2007; Mellers, 2000; Gilbert and Ebert, 2002). If possible, would it not be advantageous to learn from previous mispredictions, or barring that, listen to the advice of those who have come before?

Many reasons have been hypothesized as to why individuals do not learn from their previous affective errors (McConnell et al., 2011; Gilbert et al., 2004; Hoerger et al., 2009; Hsee and Zhang, 2004; Kahneman and Sugden, 2005). Most of these presume that the impact bias is detrimental and caused by failures in human cognition, including biased memories (Morewedge et al., 2005; Fredrickson et al., 1993), overemphasis of one's current focus (focalism bias, Schkade and Kahneman, 1998; Strack et al., 1988; Gilbert et al., 1998; Kahneman et al., 2006), and limitations in imagining future events (Kushlev and Dunn, 2012; Gilbert and Wilson, 2009). These account for the bias' proximate mechanism, but not how the bias ultimately survived natural selection. Surprisingly, few explanations have been put forth for why the impact bias might be beneficial (Miloyan and Suddendorf, 2015; Marroquín et al., 2013).

Here, we present an ultimate explanation for the impact bias. We demonstrate that a noisy environment

is the only necessary requirement to generate selective pressure for the impact bias. We extend the research on Error Management Theory (Johnson et al., 2013; Haselton and Nettle, 2006), which argues that evolution minimizes expected overall costs as opposed to the quantity of errors (McKay et al., 2009). We show that in noisy environments it can be optimal to exaggerate affective forecasts. Next, we demonstrate that when there is a benefit in exaggerating affective forecasts, an individual may perform worse if they attempt to update their forecasts from previous affective experiences or listen to accurate social advice. Finally, we discuss how this model explains more experimental evidence than previous models (see General Discussion).

4.3 Background

4.3.1 The Motivational Force of the Impact Bias

Before analysing whether the impact bias plays an adaptive role, we should first ask whether it plays any role in moderating behaviour. If individuals bias future affective experiences, but do not act on the inaccurate predictions, then the impact bias is innocuous. Interestingly, evidence not only points to the motivating power of affective forecasts, but also that the impact bias may actually lead to healthier choices.

Individuals put more effort into achieving tasks as their affective forecasts rise (Morewedge and Buechel, 2013; Greitemeyer, 2009). When given an impossible task, persistence in the futile act is linked to higher affective forecasts (Greitemeyer et al., 2011). Further, inaccurate affective forecasts may increase motivation; when participants are asked to choose from one of two tasks, the impact bias increases for the task they select (Morewedge and Buechel, 2013).

Interestingly, the impact bias may actually motivate healthy behaviour. Reduced motivation to engage in the world is inversely correlated to the impact bias. Accurately predicting the enjoyment of future affective experiences (i.e. reduced impact bias) is correlated to suicide attempts and escape fantasies (Marroquín et al., 2013). Further, when activities are of negative valence, the impact bias appears to inhibit action. If an individual exaggeratively predicts the negative consequences of an action, they are less likely to perform that action. Problem gamblers do not exhibit an impact bias, they accurately predict how poorly they feel after a loss (Willner-Reid et al., 2012).

Importantly, and despite this evidence that the bias leads to healthier behaviour, the impact bias can prove harmful. Individuals predict exercise will be less enjoyable than it is — reducing the impact bias increases intention to exercise (Ruby et al., 2011). One of the most prominent reasons for not completing a colonoscopy is fear of the pain (Janz et al., 2007; Dillard et al., 2010). Obviously, it is detrimental to remain ignorant of cancer because of exaggerated predictions of pain. Further, if an individual knows a situation is doomed to failure, larger affective forecasts are associated with fruitlessly persisting in an action regardless (Greitemeyer et al., 2011). Given that the impact bias appears to motivate at least some undesirable behaviour, it becomes imperative to elucidate whether and when the impact bias garners a benefit, and when it might be better excised.

4.3.2 Robustness Against Learning

The impact bias would be less surprising if we fixed it by learning from our mistakes. However, the bias seems robust against experiential learning (Meyvis et al., 2010; Lacey et al., 2006; Scheibe et al., 2011; Wilson et al., 2003b). In general, we are ignorant of our poor predictions, making it difficult to learn from the mistakes we do not realize we are making (Kahneman, 2011). When predicting the enjoyment of jelly beans, individuals consistently do not update their predictions even after a failed trial. Only when participants are informed of their error will they adjust their predictions (Novemsky and Ratner, 2003). However, they still will not generalize the lesson to other biased assessments, e.g. from jelly beans to music. Meyvis et al. (2010) demonstrated in a variety of contexts that individuals misremember the accuracy of their forecasts, believing that they predicted more accurately than was objectively the case. This faulty memory made it more difficult to learn from previous mistakes and correlated to persistence in the impact bias (Meyvis et al., 2010).

Not only do we not learn from our experience, but we do not listen to others. Humanity's use of social information is pervasive, lending implicit credence to its evolutionary importance (Boyd and Richerson, 1985; Rendell et al., 2010). The utility of social information is clear, it can be quickly passed between conspecifics where acquiring information through personal experience is, at worst, dangerous and, at best, time consuming (Fernández-Juricic and Kacelnik, 2004; King and Cowlishaw, 2007; Magurran and Higham, 1988). In fact, humans seem biased toward employing social information, even when it is incorrect (Whiten et al., 2009; Sommerfeld et al., 2008; Rauwolf et al., 2015). What is interesting, however, is that while in many contexts individuals are so biased toward social information that they apply *inaccurate* social information, people often refuse to utilize *accurate* social information in the realm of affective forecasting.

In Gilbert et al. (2009) participants were tasked with predicting their enjoyment of a speed date. To aid them, they were offered a choice of information: (i) a dossier of the prospective date, or (ii) the affective experience of another individual who had already met the prospective date. The majority requested a dossier, despite the fact that once the participant went on the date, the original social advice was a significantly better predictor of their own experience compared to their prediction (Gilbert et al., 2009). Furthermore, when individuals are given both social advice and information about an impending experience, they still significantly bias their predictions toward their own inaccurate personal beliefs (Guarnaccia, 2012). If humans are typically predisposed to using social information, why, in the realm of affective forecasting, would we ignore accurate social information in lieu of consistently biased personal predictions?

4.3.3 Proximate Explanations

Significant research has gone into the proximate mechanisms which maintain the impact bias (McConnell et al., 2011; Gilbert and Wilson, 2009; Kahneman et al., 1997; Gilbert et al., 2004; Hoerger et al., 2009; Hsee and Zhang, 2004; Kahneman and Sugden, 2005). Because the bias appears to be generated by mechanisms such as myopic focus (Schkade and Kahneman, 1998; Strack et al., 1988; Gilbert et al., 1998; Kahneman

et al., 2006) and biased memories (Morewedge et al., 2005; Fredrickson et al., 1993), it has been easy to assume that impact bias generates suboptimal decisions and, if possible, should be regulated (Feys and Anseel, 2015; Peters et al., 2014). However, if there is utility in the impact bias, then the mechanisms which evolved to create and maintain the bias may look precisely like biased memories and/or a tendency toward myopic focus.

Individuals demonstrate the impact bias when they myopically focus on the consequences of a decision. For example, when predicting the happiness of a football team winning, individuals forget about the other aspects of their life, such as laundry; consequently they exaggerate the effect of a football victory (Wilson et al., 2000). Reminding people of the other factors in their life reduces or eliminates the bias (Schkade and Kahneman, 1998; Buehler and McFarland, 2001; Hoerger et al., 2010, though see Sevdalis and Harvey, 2009). If individuals keep a journal about their daily life, then their affective forecasts are less biased (Wilson et al., 2000).

Proximate reasons have also been hypothesized as to why we do not learn to improve our predictions over time. One such hypothesis attributes culpability to memory biases. Recollection of past events has been found to be biased toward atypical (Morewedge et al., 2005) and recent (Fredrickson et al., 1993) experiences. If we remember experiences incorrectly, then it is difficult to correct forecasting errors (Kahneman et al., 1997). Furthermore, when we imagine a future state, we employ the memories that are most available, rather than the most typical (Gilbert and Wilson, 2009). Thus, if we formulate predictions based on the most memorable experiences, and those memories are biased toward outlying experiences, then it is unsurprising that we exaggerate our predictions.

Whilst such research elucidates mechanisms which generate the impact bias (proximate mechanisms), it does not speak to how the bias survived evolution. While the impact bias can be attenuated by reminding individuals of mundane activity, such research does not consider when culling the bias is best. As previously mentioned, the impact bias is linked to both healthy (Marroquín et al., 2013; Willner-Reid et al., 2012) and harmful (Ruby et al., 2011; Janz et al., 2007; Dillard et al., 2010) behaviours. Thus, it may be dangerous to presume *a priori* that the impact bias is detrimental. The important open question is whether the bias provides any overall functional utility.

4.3.4 Does the Impact Bias Provide a Selective Advantage?

Given the pervasiveness and motivational power of the impact bias, surprisingly little research has considered whether it performs a function (Miloyan and Suddendorf, 2015; Marroquín et al., 2013). To our knowledge, Robson and Samuelson (2011) offer the only evolutionary model which suggests an explanation for why the impact bias may have survived natural selection. They do this by extending a class of models which attempt to explain hedonic adaptation (Perez-Truglia, 2012; Graham and Oswald, 2010; Rayo and Becker, 2007).

Hedonic adaptation is the propensity for feelings to regress to the mean over time (Frederick and Loewenstein, 1999). For example, a person is about as happy two years after winning the lottery as they were prior

to the monetary windfall (Brickman et al., 1978). Hedonic adaptation is fairly pervasive, though context appears to significantly moderate whether someone adapts to life events. For instance, individuals quickly adapt to pay raises (Tella et al., 2010) and widowhood (Clark et al., 2008), but do not adapt to unemployment (Clark et al., 2008) or traffic noise (Weinstein, 1982).

One of the reasons individuals appear to inaccurately predict their affective experiences is because they do not compensate for hedonic adaptation when forecasting (Gilbert et al., 1998; Ubel et al., 2005; Nelson and Meyvis, 2008). Related, there is evidence that the intensity of the impact bias is correlated to the success of an individual's coping mechanism (Gilbert et al., 2004; Hoerger et al., 2009). Because individuals do not compensate for how well they will cope, improved coping mechanisms lead to larger impact biases.

Robson and Samuelson (2011) argue that this ignorance of hedonic adaptation could lead to selective pressure for the impact bias. If there are physiological limitations on (i) the intensity of feelings, and (ii) our ability to differentiate between goal states with similar utility, then ignorance of future experiences can be adaptive. The intuition of the argument presented by Robson and Samuelson (2011) is that individuals exhibiting the impact bias are correctly predicting the utility of an action, but then, after taking an action, the individual's affective experience quickly adapts to another goal.

However, there is a potential problem with this explanation. Though coping and ignorance of hedonic adaptation are correlated with the impact bias, both mechanisms take time, whilst the impact bias appears immediately after an event (Wilson and Gilbert, 2013, though see Levine et al., 2012). In Gilbert et al. (1998) the ability to cope was experimentally manipulated and affective forecasts were measured against affective experiences both (i) immediately after the event, and (ii) after ten minutes. Ten minutes after an event, those who easily coped had hedonically adapted more than the difficult coping group. Because neither group had predicted they would cope after ten minutes, the difference between forecasts and experiences (i.e. the impact bias) was larger in the coping group. However, immediately following the event, there was no measurable difference between the affective experiences of the groups — coping and hedonic adaptation had yet to take place. Interestingly, the impact bias was still clearly present. Regardless the coping mechanism, immediately following an event, individuals still incorrectly predicted their affective experience (Wilson and Gilbert, 2013).

Whilst Robson and Samuelson (2011) have gone some distance in explaining the function of the impact bias, there is still significant work to be done. Though hedonic adaptation and coping may explain some of the experimental results, it does not explain why the impact bias exists immediately following an affective experience. Further, there are additional individual differences regarding the impact bias which are, as yet, unexplained.

4.3.5 Individual Differences

Individual difference data for the impact bias is fairly nascent (Wenze et al., 2013). However, any attempt to explain the function of the impact bias should attempt to account for the individual differences. Beyond

coping mechanisms, two factors appear linked to the size of the impact bias: (i) perceived importance of the event, and (ii) Emotional Intelligence.

The size of the impact bias is correlated to the perceived importance of an event; the more intense the expected feeling, the larger the error between affective forecast and experience. The more enjoyable a romantic relationship, the larger the impact bias when considering the pain of a break-up (Eastwick et al., 2008). Further, the perceived importance of an election is correlated to the size of the bias (Hoerger et al., 2010). This tendency holds even when an event's importance is experimentally manipulated (Greitemeyer, 2009).

The second factor is that Emotional Intelligence is inversely correlated to the impact bias. When individuals are better able to sense their own emotions, they exhibit less of an impact bias (Hoerger et al., 2012). Dunn et al. (2007) showed that the impact bias is inversely correlated to several features of emotional intelligence, such as perception of emotion, use of emotion to facilitate thought, understanding of emotion, and management of emotion. Further, understanding the interplay between emotions and exogenous events, an aspect of mindfulness, is linked to impact bias (Emanuel et al., 2010).

In this article we not only attempt to explain the evolutionary benefit of the impact bias, but also why the bias is correlated to the intensity of the event and inversely correlated to Emotional Intelligence. To do this, we extend the literature on Error Management Theory. We argue that the impact bias is adaptive when decision-making in noisy environments.

4.3.6 Error Management Theory

Error Management Theory (EMT) predicts that in noisy environments, where different actions lead to asymmetrical benefits/costs, biasing information when making a decision can be optimal (Johnson et al., 2013; Buss and Kenrick, 1998). The underlying idea is well explained by an example from McKay et al. (2009). If building a perfect smoke detector is impossible, it is better to error on the side of caution. Sounding an alarm when there is no fire is annoying, but not sounding the alarm when there is a fire is deadly. Thus, it is advantageous to bias the detector toward hyper-sensitivity. Even though this results in more errors, it results in paying a lower cost — less death, but more annoyance.

EMT has predominately been used to argue that, in unpredictable environments, it is advantageous to bias one's belief in the probability that an event will occur. In general, humans tend to bias their predictions about the likelihood of an experience, predicting they are more likely to experience positive events, and less likely to face negative events compared to the average person (Alicke et al., 1995). We believe that compared to others we are better drivers (DeJoy, 1989; Dalziel and Job, 1997), have a reduced chance of divorce (Weinstein, 1980), and will live longer (Weinstein, 1980, see Sharot et al., 2007 for a nice review). EMT demonstrates that in an unpredictable world such biases can be optimal by compensating for noisy environments (Haselton and Nettle, 2006; McKay et al., 2009; Sharot, 2011). Further, EMT has been used to described how biased probabilistic thinking may explain human behaviour related to anxiety (Bateson

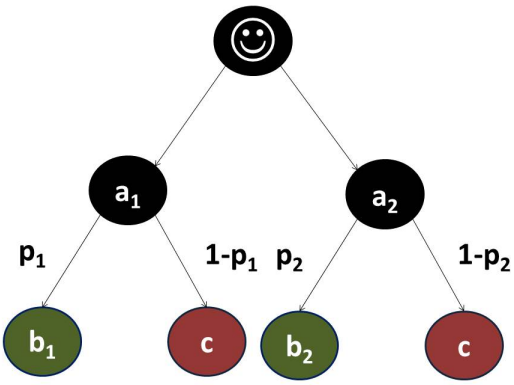


Figure 4-1: The agent chooses a_1 or a_2 and with some probability (p_1 or p_2) succeeds and receives a benefit (b_1 or b_2) in the range $[-1, 1]$. If it fails, it pays a cost c.

et al., 2011), belief in god (Johnson, 2009), perceived sexual attraction (Haselton and Buss, 2000), and social ostracism of potentially diseased individuals (Kurzban and Leary, 2001).

In this paper, we consider that Error Management Theory may not only explain why people exaggerate their belief in the likelihood of an event's occurrence (e.g. the optimism bias), but also why we exaggerate the predicted reward for an action (i.e. the impact bias). While EMT shows that it is beneficial to exaggerate information when making a decision under noise, it is agnostic on whether it is best to exaggerate the likelihood of an event, or to exaggerate one's belief in an action's reward. Exaggerating the probability that an event will occur should motivate action, but so should exaggerating one's belief in an action's reward.

There is some debate, however, regarding whether EMT leads to cognitive biases such as optimism and the impact bias (McKay and Efferson, 2010). Trimmer et al. (2011) analysed whether it is advantageous to add bias when attempting to learn the probability of an event occurring. They showed that it depends on the algorithm exercised when attempting to learn the probability. However, such research focuses on attempting to learn a static probability over time. In nature, the probability of an event occurring may itself be probabilistic. An event may occur with probability 0.3 in one situation, but then with 0.4 in another. The presumption of using static probabilities for evolutionary models on decision-making has recently come under criticism (Fawcett et al., 2014).

In this paper we rectify this constraint, whilst simultaneously side-stepping the fact that biases are not necessarily advantageous whilst learning static probabilities (Trimmer et al., 2011). In our model, in every round there is a new probability generated for the chance of succeeding at a task. In this sense, the environment is not predictable, and arguably more natural.

In the next two sections we extend the research on EMT to the impact bias. We analyse how the impact bias could prove adaptive in noisy environments. Additionally, we demonstrate why it may be best not to learn from past affective experiences. Finally, we show that the predictions of this model match the

individual difference data, lending credence to the hypothesis that the impact bias may have evolved at least in part to augment decision-making under noise.

4.4 Study 1: Optimal Decisions in Noisy Environments

4.4.1 Model 1a: Predictable Environments

An agent-based, evolutionary approach is used to analyse whether the impact bias can be adaptive. We simulate a situation where an individual must decide between two actions by predicting the benefit of the actions. We then analyse whether there are contexts where exaggerating affective forecasts is advantageous.

We begin with a population of agents where each agent plays multiple rounds of a game against nature. During each round, an agent must choose between two actions, action 1 (a_1) or action 2 (a_2). After selecting an action the agent either succeeds or fails at the task. Prior to choosing, each agent is given the probability of accomplishing the chosen action (p_1 and p_2 in the range $[0, 1]$). If the agent chooses an action and succeeds, then some benefit (b_1 or b_2 in the range $[-1, 1]$) is added to the agent's fitness value. Otherwise, it pays a heavy cost, $c = -2$ (see Figure 4-1).

The information employed in making the decision is referred to as one's *decision utility* (Kahneman et al., 1997). The optimal decision utility leads one to choose a_1 when:

$$p_1 b_1 + (1 - p_1)c \geq p_2 b_2 + (1 - p_2)c \tag{4.1}$$

For each game, the agent is given p_1, p_2, c, and b_2, but *not* b_1. Each agent consists of a single gene representing its belief in the benefit of b_1, represented as b_1'. Since the actual benefit of succeeding at a_1 is hidden from the agent, each agent employs its belief (b_1') in b_1 as part of its decision utility. Namely, the agent selects a_1 if:

$$p_1 b_1' + (1 - p_1)c \geq p_2 b_2 + (1 - p_2)c \tag{4.2}$$

Each round a new p_1, p_2, and b_2 are selected. b_1 and c remain constant across games. The population consists of $N = 3000$ agents and in each of $g = 2000$ generations every agent plays $r = 3000$ rounds of the game. Initially, each agent's belief (b_1') is chosen randomly in the range $[-1, 1]$. At the end of each generation, agents are selected for the next generation based on their relative fitness (Zitzler and Thiele, 1998). Each selected agent's allele (b_1') plus a Gaussian mutation rate ($\mu_{mut} = 0$ and $\sigma_{mut} = 0.01$) is used for the next generation. Though the agents' initial beliefs are restricted to the range $[-1, 1]$, they may mutate outside that range. We employ an evolutionary algorithm here not because we believe affective predictions are stored genetically, but rather as an algorithm which searches for the optimal belief.

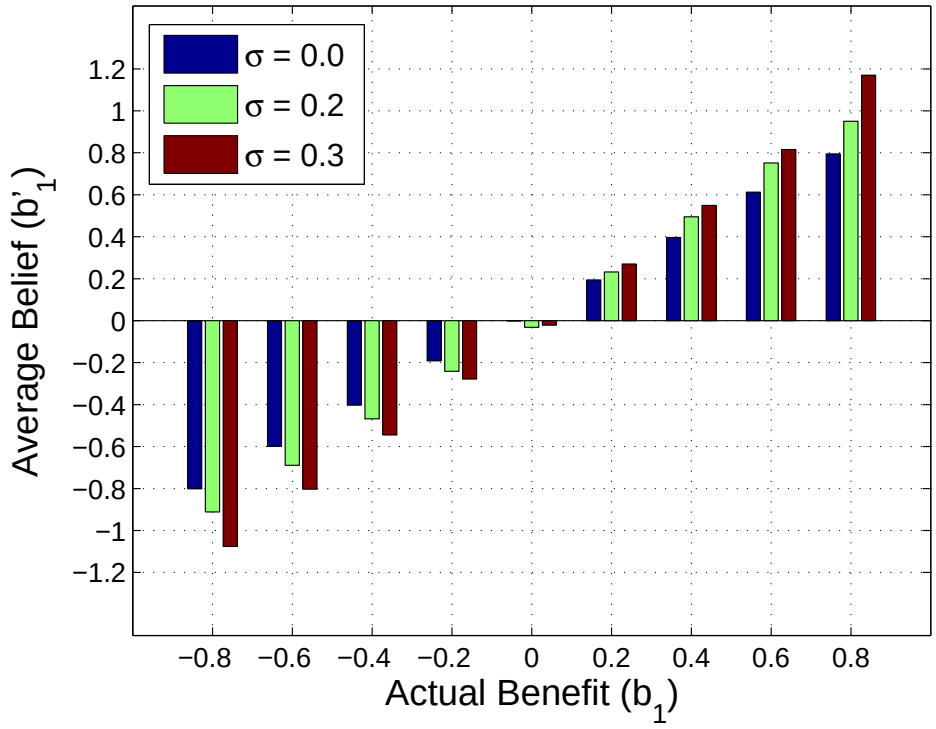


Figure 4-2: Average value of b'_1 over the last 300 generations. Each colour represents the variance of the noise added to p_1 and p_2. **Left bar:** no noise (model 1a). **Center bar:** $\sigma = 0.2$, and **Right bar:** $\sigma = 0.3$ (model 1b).

Results

The model was run for different values of b_1. The blue (left) bar in Figure 4-2 shows the population's average value of b'_1 after g generations. As expected, the agents perform best when $b'_1 = b_1$. This is because the agent's decision utility (equation 4.2) equals the optimal decision utility (equation 4.1). Since b_1 is the reward the agent receives if it succeeds at task a_1, b_1 represents the actual affective experience of the agent. The agent performs optimally when its affective forecast b'_1 correctly predicts the reward the agent will experience, b_1. There is no evidence of impact bias.

4.4.2 Model 1b: Unpredictable Environments

Environments are often not fully predictable. To model this, noise (ϵ) is added to the probabilities of success (p_1 and p_2). The precise values of p_1 and p_2 are now obfuscated from the agents. Each agent is given p'_i, where $p'_i = p_i + \epsilon_i$. ϵ is randomly selected per round over a Gaussian distribution with $\mu = 0$ and some

constant σ.

An agent selects a_1 when:

$$p_1' b_1' + (1 - p_1')c \geq p_2' b_2 + (1 - p_2')c \tag{4.3}$$

Results

Figure 4-2 depicts the population's average value of b_1' when the variance (σ) of noise (ϵ) added to the probabilities is held at 0, 0.2, and 0.3. Model 1a represents a predictable environment where $\epsilon = 0$ and the agents perform optimally when their prediction of an experience matches the actual experience, (i.e. $b_1' = b_1$). When noise is added to p_1 and p_2 the optimal belief (b_1') is no longer the actual benefit b_1. As b_1 drops below zero, it is beneficial for $b_1' < b_1$. As b_1 rises above zero, it is beneficial for $b_1' > b_1$.

Two additional results are of note. First, the intensity of the actual emotion ($|b_1|$) is correlated to the size of the optimal bias ($|b_1' - b_1|$). Figure 4-2 shows that as $|b_1|$ increases, $|b_1' - b_1|$ increases. Second, the amount of noise is correlated to the size of the bias. When $\sigma = 0.3$, the optimal bias is larger than when $\sigma = 0.2$. The implications of these findings are analysed in the General Discussion.

Discussion

The impact bias is defined by exaggeratively predicting both positive and negative affective experiences. Individuals predict they will be happier for positive events and less happy for negative events compared to their eventual experience. Study 1 demonstrates that, in noisy environments, agents employing an appropriate impact bias would outperform agents without an impact bias. When making a decision and forecasting the benefit of an action, agents developed overly positive beliefs for positive events ($b_1' > b_1 > 0$) and exaggeratively negative beliefs for negative events ($b_1' < b_1 < 0$). There is, thus, selective pressure for an impact bias. This result offers an explanation for recent work demonstrating that the impact bias increases motivation (Morewedge and Buechel, 2013; Kwong et al., 2013). Motivating a decision by exaggerating affective forecasts can be advantageous in noisy environments.

This result extends the literature on Error Management Theory to the impact bias. Noise introduces error into the decision-making process. To compensate, the agent biases its decisions to the, on average, larger benefit. When $b_1 > 0$, the benefit will, on average, be larger than b_2, thus the agent selects b_1 more frequently by biasing its belief of b_1 higher. When $b_1 < 0$, it biases its decisions toward b_2, reducing its belief of b_1.

4.5 Study 2: Learning (or not) from Personal Experience

4.5.1 Model

We now ask whether it is beneficial to learn from experience. We assume an agent's affective experience evolved to match the fitness benefits of the experience. So, when an agent receives b_1, it affectively experiences b_1.

The agent is given repeated exposure to b_1, which it stores in memory. When it is deciding which action to select, it can use either its evolved belief, b'_1, or knowledge of its past experience, b_1. If it uses its evolved belief, then it makes a decision based on Equation 4.3. If it uses its experience, then it replaces b'_1 with b_1 in Equation 4.3.

The conditions from Study 1 remain the same, except we add an additional gene, α to each agent. Here, α represents the percent chance an agent uses its affective experience (b_1) instead of its evolved belief (b'_1) for each round of the game. Each agent's genome now consists of a belief in the benefit of a_1 (i.e. b'_1), and α. When an agent is selected for the next generation, both traits are passed on and subject to Gaussian mutation with $\mu_{mut} = 0$ and $\sigma_{mut} = 0.01$. The initial agents' α are randomly chosen in the range $[0, 1]$. If α mutates outside that range it is normalized to the nearest limit when applied to the decision equation.

Results

Figure 4-3 illustrates the population's mean value of α in environments with different noise rates when $b_1 = -0.8$. The value $b_1 = -0.8$ is chosen for clarity; the results hold wherever the impact bias is advantageous. Without noise (circles) α is positively selected — direct experience is favoured. With noise (triangles and squares), α drops below zero. Agents who do not learn from their experience outperform those who do.

Discussion

In noiseless environments, it is optimal for $b'_1 = b_1$ (Study 1a). It is not advantageous to possess an impact bias, and thus, there is also selective pressure to learn from experience (Figure 4-3: circles). In noisy environments, it is optimal to exaggerate one's belief in the actual benefit of the experience ($|b'_1| > |b_1|$) — there is an advantage in possessing the impact bias (Study 1b). Consequently, when agents are given the opportunity to eliminate their bias and learn from experience, evolution rejects this offer, employing the impact bias and ignoring personal experiences (Figure 4-3: triangles and squares). This offers a potential explanation as to why humans appear not to learn from their affective mispredictions (Meyvis et al., 2010; Lacey et al., 2006; Scheibe et al., 2011; Novemsky and Ratner, 2003).

Of note, we are not claiming that humans evolved a belief for every action. This model is an existence proof that in noisy environments it is advantageous to evolve some kind of bias. Furthermore, we are not claiming that individuals do not learn from their affective experience. Individuals may learn from their experience, but then subsequently add the impact bias for use during decision-making. Affective experience

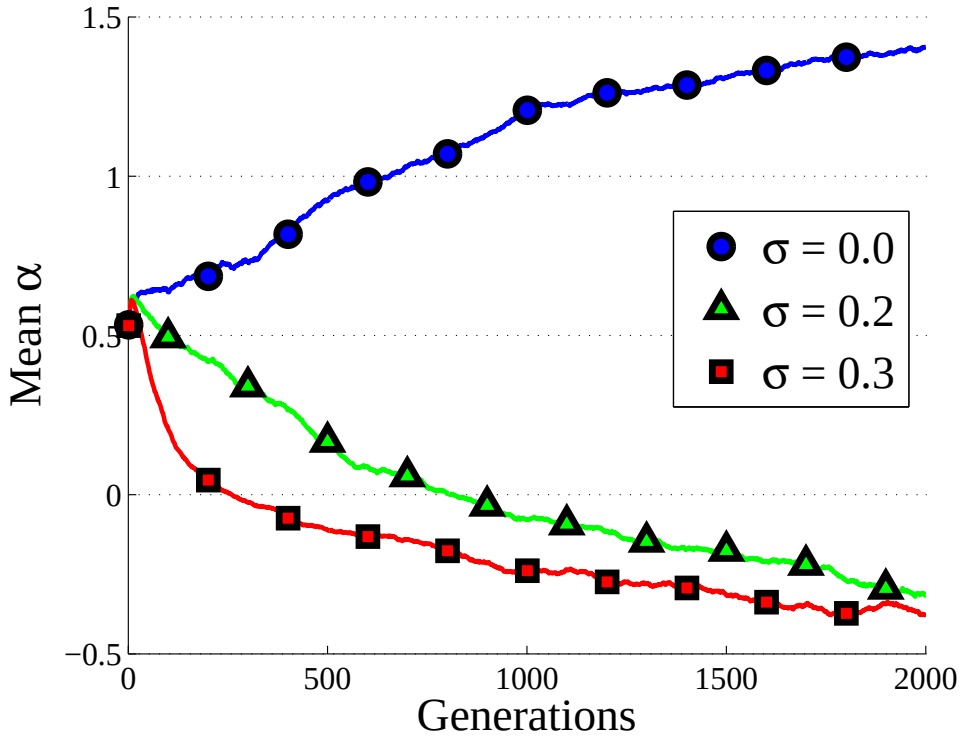


Figure 4-3: The population's mean α in noisy and non-noisy environments, averaged over 10 runs. **Circle:** Without noise, $\sigma = 0.0$ **Triangle:** With noise, $\sigma = 0.2$. **Square:** $\sigma = 0.3$

likely still grounds decision utility. In our simple model the agents could not alter their experiences with a bias, so they disregarded them. If humans (unconsciously) update their experience with a bias, our model posits why individuals *appear* not to learn from experience. For example, someone may learn that on a scale of 1-10 a romantic break-up feels like a 3. However, if the impact bias is advantageous, they may unknowingly bias their predictions of a break-up to a 1, optimizing decisions. As a consequence it will appear as if they did not learn.

Finally, this result offers an explanation regarding why people ignore accurate social advice (Gilbert et al., 2009; Guarnaccia, 2012). In the model an individual is given information on the benefit of an action (b_1), and must decide whether to use the information, or rely on its biased belief (b_1'). The information about b_1 could come from first hand experience, but it could also come from someone else's experience. As shown in Gilbert et al. (2009), an individual's actual experience of an event (b_1) and another's person's experience of an event are quite similar. If we replaced personal knowledge of an event with another person's advice, the results of the model will still hold. While two people's experience of an event would not be identical, if they are close enough, then social advice should be ignored for the same reason that personal experience should

be ignored — it is advantageous to maintain the impact bias. This may explain why, despite humanity's pervasive use of social information, we ignore it when performing affective forecasting.

4.6 General Discussion

We have presented an ultimate evolutionary explanation for the impact bias and its robustness against both experiential and social learning. In noisy environments it can be advantageous to exaggerate one's affective forecast relative to one's affective experience, and consequentially beneficial not to (i) learn from past experiences, or (ii) listen to social advice. Next, we show that this model offers insight into previously unexplained individual difference data, adding credence to the hypothesis that the impact bias evolved in humans, at least in part, to optimize decision-making in noisy conditions.

4.6.1 Intensity Increases Impact Bias

Figure 4-2 shows that as the intensity of the actual benefit ($|b_1|$) increases, so does the size of the optimal impact bias ($|b_1' - b_1|$). If our hypothesis is correct and the impact bias evolved for humans to navigate noisy environments, we would expect this correlation to appear in nature. It seems to. During the termination of romantic relationships, the more one enjoyed the relationship, the larger the forecasting error (Eastwick et al., 2008). When predicting one's emotion following a presidential election, the size of the impact bias is correlated with the perceived importance of the election (Hoerger et al., 2010). Even when the importance of an event is experimentally manipulated, those who feel more intense emotions during an experience predict more inaccurately prior to the experience (Greitemeyer, 2009).

4.6.2 More Noise Increases Impact Bias

Figure 4-2 shows that as the amount of noise increases, the optimal bias increases. Again, if the impact bias evolved for humans to navigate noisy environments, we would expect to witness this phenomenon. We do. There is evidence that the impact bias is reduced when an individual can sense themselves better. Accuracy in affective forecasts is correlated to emotional perception, an aspect of Emotional Intelligence (Hoerger et al., 2012). Someone who more accurately perceives their emotions, also more accurately predicts their emotions. In fact, the impact bias is inversely correlated to several vectors of Emotional Intelligence (Dunn et al., 2007). Related, "knowledge of the interplay between internal emotions and external events", an aspect of mindfulness, is inversely correlated to an individual's impact bias (Emanuel et al., 2010).

Our model is agnostic on whether noise originates from the unpredictability of the environment or from limitations in the agent's ability to perceive their own responses. If individuals are better able to predict the environment (including their own reactions), then there is less need for a compensating bias. In contrast, if someone is not adept at sensing their emotions, their affective experience is a noisy signal of their actual emotion, and the impact bias mitigates the negative repercussions of the noise.

4.6.3 Probabilistic Information Alters Affective Forecasts

Recently, researchers have begun to investigate how knowledge of the likelihood of an event's occurrence alters affective forecasts. Affective forecasts differ based on whether there is a 90% or 10% chance of receiving a reward (Buechel et al., 2014). When preparing to play a competitive game, participants who know for certain their opponent's identity, display a larger impact bias when predicting their joy at victory (Morewedge and Buechel, 2013). Why would changes in likelihood of events affect affective forecasts?

At first glance, it seems to make little sense that the value of a reward changes based on the likelihood of attaining that reward. Certainly the motivation to pursue the reward should change based on the likelihood of success, but not the value itself. For instance, if an individual is given the opportunity to receive an apple, then, at that moment, that apple is worth some value, let us say X. If there is a .1 or .9 chance of attaining the apple, that should not impact the forecasted value of the apple; it should remain X. However, one may not wish to pursue the apple if there is a low probability of success, because the value of the apple does not warrant the risk. The current model demonstrates that if the probability of attaining the apple is noisy, it can be advantageous to bias the forecasted value of the apple; i.e. the value of the apple is X + bias. However different probabilities should not alter the perceived value.

Invariably there are several factors which generate the impact bias. We have presented a theory which coalesces hitherto unexplained experimental data. However, we do not explain why different probabilities of success lead to changes in affective forecasts. Perhaps, as Morewedge and Buechel (2013) hypothesize, probabilistic information affects the motivation of the individual, causing feedback into affective forecasts. It would be interesting to see further research integrating this research with our theory. Open questions remain, such as how a mixture of noisy and 'clean' probabilistic information affects the impact bias.

4.6.4 Impact Bias In Noise Free Situations

A potential criticism of the model is that the impact bias persists in studies where information is certain. If the impact bias evolved to navigate noisy environments, why would it persist in predictable contexts?

We accept this criticism but suggest that in the natural world individuals rarely have the luxury of certain information. As discussed above, even when the outside world is predictable, an understanding of the self is imperfect and noisy. Something as simple as a bad night sleep can impact cognition (Gruber et al., 2010). Furthermore, there is evidence that even when information is fairly certain, we do not process it accurately.

Evidence suggests that individuals are insensitive to probabilistic variations when making decisions about affective outcomes (Loewenstein et al., 2001). Rottenstreich and Hsee (2001) show that humans process a 99% and a 1% chance of receiving a shock quite similarly. When choices are affect-rich, individuals are less sensitive to probability (Pachur et al., 2014). When predicting future affective experiences, increasing affective intensity diminishes sensitivity to probability (Buechel et al., 2014). Suter et al. (2015) hypothesize there is a different underlying cognitive decision-making mechanism for affect-rich and affect-poor contexts. In their study, the best model for predicting human decisions in affect-rich contexts includes

ignoring probabilistic information. Since probabilistic insensitivity is another form of noise, the impact bias could be advantageous.

Importantly, however, we are not claiming that impact bias is always beneficial. As previously mentioned, the impact bias can lead to suboptimal decisions (Ruby et al., 2011; Janz et al., 2007; Dillard et al., 2010). Here we showed that, in noisy environments, the bias can be functional. If the impact bias is employed in a situation where information is transparent and known, we would expect the bias to lead to suboptimal decisions. We hope this work advances the conversation regarding when and how the impact bias should be regulated therapeutically.

4.6.5 Self-Deception

Why are we generally ignorant of our poor affective forecasting performance (Kahneman, 2011)? Even if biasing affective forecasts is optimal, why not be aware of that fact and knowingly bias our predictions? We believe this question may relate to the theory and evolutionary utility of self-deception (von Hippel and Trivers, 2011a; Trivers, 1991; McKay et al., 2009; Smith, 2014; Mele, 2001; Trivers, 2011). The impact bias appears to exemplify a situation where one's behaviour and one's cognitive understanding of one's behaviour appear to diverge — which is self-deceptive.

Recently, it was shown that self-deception can be adaptive in order to compensate for noisy information (Ramirez and Marshall, 2015). The underlying mechanism in Ramirez and Marshall (2015) is very similar to the one presented here. They show that when deciding whether to fight over a resource, if knowledge about the probability of success is noisy, it may prove advantageous to believe you are stronger than you are. Importantly, however, even if using biased information is beneficial, believing that the biased information is true is not necessary (McKay and Efferson, 2010). In Ramirez and Marshall (2015), an agent's fitness increased when it believed it was stronger than was objectively the case. However, the agent would perform just as well if it accurately predicted its chance of winning, but also knew it was optimal to *act* as if it was stronger than was the case (Frankish, 2009).

Similarly with the impact bias, while the optimal behaviour is to maintain the bias, this can be done in one of two ways. A person can inaccurately forecast and remain ignorant of the error; this is what we see in the experimental data. Given this ignorance, they never learn to update their behaviour and the bias is maintained. In this case, the individual is self-deceived. However, a person would perform just as well if they accurately forecast their affective experiences, but also knew it was advantageous to bias the predictions when making a decision. In this case, the individual would not be self-deceived. They would know that biasing affective forecasts is beneficial, so they would not update their forecasts from experience.

When would evolution choose self-deception over an accurate understanding of the world? Typically, the debate has centered around whether it is cheaper or easier to evolve self-deception compared to a fully accurate understanding of the world (McKay and Dennett, 2009). As such the evolutionary pressures and constraints which led to the development of the mind need to considered more frequently in evolutionary

modelling (McNamara and Houston, 2009). Unfortunately, to date, too little is known to predict the evolution of cognitive biases like self-deception (Haselton and Buss, 2009; Marshall et al., 2013c). However, given that individuals are consistently unaware of their forecasting errors (Kahneman, 2011; Novemsky and Ratner, 2003), and that, despite our pervasive use of social information, we tend to ignore accurate social advice when forecasting (Gilbert et al., 2009; Guarnaccia, 2012), it may be time to take serious the idea that self-deception is playing a role in human evolution.

4.6.6 Conclusion

We employed an evolutionary model and posited an ultimate evolutionary explanation for the impact bias and why individuals appear not to update their affective forecasts with experiential or social learning. We show that in noisy environments agents using the impact bias outperform those that do not. We then demonstrate that an agent's success may further hinge on not learning from its affective experience, or applying accurate social advice.

Few previous models posit a functional explanation of the impact bias (Miloyan and Suddendorf, 2015; Marroquín et al., 2013). Those that have focus on people's tendency to hedonically adapt over time (Robson and Samuelson, 2011). Our explanation goes beyond previous work by explaining why the impact bias is present immediately after an experience.

Our model accounts for previously unexplained experimental results, namely that both emotional intensity and increased noise amplify the bias. We believe this work and the further analysis of the existing empirical evidence demonstrate that the impact bias may well serve an important function. Any explorations regarding whether, when, or how to address the bias in patients should consider its potential important role in decision making.

5

Trust Mediates Costly Punishment Given Partial Information and Partner Choice

> *Those who see me rarely trust my word: I must look too intelligent to keep it.*

Jean-Paul Sartre, *The Devil and the Good Lord*

> *I have made it a rule of my life to trust a man long after other people gave him up, but I don't see how I can ever trust any human being again.*

Ulysses S. Grant,

5.1 Summary

Blind trust in others and costly punishment of those who refuse to treat others fairly are each independently essential to human sociality. Yet neither trust nor punishment have proven easily explicable through standard evolutionary theory. By analysing the Trust Game, we show here that both phenomena can be adaptive in the context of partial information and partner choice. Blind trust develops based on prior experience of a society's trustworthiness, and in turn mediates the evolutionary feasibility of costly punishment by generating competition between known and unknown offers. Costly punishment through the rejection of advantageous offers known to be unfair then becomes adaptive, creating a context that generates unstable yet

highly reliable cooperation. This result is important because it demonstrates that the evolutionary viability of trust, fairness, and costly punishment may be linked given only the plausible assumptions of partial information and partner choice. That fairness and cooperation can result from these may imply that norms of fairness could become more brittle in an age of increasing information transparency.

Note: This chapter does not demonstrate a context where it is adaptive to lack veridical knowledge. However, this work is the foundation for Chapter 6 and 7.

5.2 Introduction

The Trust Game offers an individual the opportunity to invest with another in hopes the other can be trusted to return an increased investment. When offered the opportunity to invest, human players tend to make two seemingly suboptimal and paradoxical decisions. First, when individuals possess no information about their partner's trustworthiness, they trust and invest despite the fact that the subgame perfect strategy is for their partner to defect (Berg et al., 1995). Second, even when participants know a trade will be profitable, they reject it unless the trade is fair — they express costly punishment against a potential benefactor (Marlowe et al., 2010).

Why would it be beneficial to reject profitable offers on the one hand, but accept unknown (and thus potentially deleterious) offers on the other? Here we demonstrate that blind trust and costly punishment can be revenue maximizing. We analyse the Trust Game in a context where investors have both partner choice and partial information about potential partners' trustworthiness. We find that trust and costly punishment can both evolve, and that trust mediates the evolution of costly punishment. Rather than representing a paradox, we find that the tendency to trust and costly punish are revenue maximizing strategies when taken in the context of each other, providing a parsimonious account for the development of fairness norms.

5.3 Costly Punishment in the Context of the Trust Game

In the Trust Game (TG), an individual (the investor) is given one unit of money. They can either keep it and walk away, or invest it with another individual (the trustee). If invested, the money is multiplied by some factor, b, and the trustee can choose to return some portion to the investor. The amount returned is determined by the trustee's return rate r; any rate greater than $1/b$ garners the investor a profit (see Table 1). In a one-shot game where the investor does not possess any information about the trustee, defection is the subgame perfect equilibrium. Because the trustee never benefits from returning anything to the investor, the investor should not trust. But humans do tend to trust, and to invest, despite its suboptimality (Berg et al., 1995). Attempts to diagnose this propensity for blind trust has spanned both empirical (Johnson and Mislin, 2011; Burks et al., 2003; Charness and Dufwenberg, 2006; Al-Ubaydli et al., 2013; Engle-Warnick and Slonim, 2004; Boero et al., 2009; Delgado et al., 2005; Kosfeld et al., 2005; Pfattheicher and Keller,

2014) and theoretical (Güth and Kliemt, 2000; Masuda and Nakamura, 2012; Bravo and Tamburino, 2008) disciplines.

Recently, the theoretical literature has investigated whether experimentally-observed levels of trust can be explained by positing investors with some knowledge of the trustee's return rate. The intuition is that if humans evolved in small groups where reputations are difficult to hide, perhaps the levels of trust observed in experiments would be adaptive. Manapat et al. (2012) have shown that selective pressure for trust can arise when the chance of knowing a trustee's return rate exceeds some threshold. This holds even when information is delayed and inconsistent (Manapat and Rand, 2012). Further, population structure and information of the trustee's return rate synergistically enable trust (Tarnita, 2015).

Introducing knowledge of individual trustworthiness into the Trust Game opens up a new way to consider human play. We may now think of the TG as it relates to the Ultimatum Game (UG), and recognise failures to trust as a form of costly punishment. In the UG, a proposer offers some fraction of a windfall to a responder. The responder decides whether to accept the offered partition, or to reject it — in which case neither side keeps any share of the investment. The subgame-perfect equilibrium is for the responder to accept any offer; but most UG responders will reject unfair offers (Marlowe et al., 2010), though what is deemed 'fair' varies somewhat across cultures (Henrich et al., 2001). In the TG, if an investor knows the expected return rate r for a trustee, their decision to invest becomes relatable to the responder's decision in the UG (Tarnita, 2015).

Recently, experimental research has considered the effects of offering investors in the Trust Game information about the trustees return rate, prior to deciding whether to invest. The investors often reject offers that would be profitable ($r > 1/b$), implying an implicit demand for fairer returns (i.e. $r \approx 1/2$) (Manapat et al., 2012). Rejecting an offer even when it provides a net gain for the investor can therefore be interpreted as a form of costly punishment.

Costly punishment is considered key to understanding the level of cooperation expressed in contemporary human society (Gintis et al., 2005; Henrich et al., 2010; Powers and Lehmann, 2013). Nevertheless, both the ultimate and proximate cause of costly punishment is still an open question (Rankin et al., 2009; Sylwester et al., 2013a; Hauser et al., 2014). Although older theoretical work seemed to indicate that the altruistic punishment of free riders might support cooperation, these models collapse in the light of the full range of empirically observed punishment, which includes the antisocial punishment of those contributing more to their society than the punisher (Rand et al., 2010; Powers et al., 2012; dos Santos, 2014). Though normally only defined in the context of the public goods game, the rejection of profitable but unfair offers in the trust game might also be seen as a form of anti-social punishment, since it punishes a benefit of greater value than the original investment (Sylwester et al., 2013b; Herrmann et al., 2008a).

Some researchers have suggested that costly punishment is explained by a human predisposition toward fairness—that individuals are willing to pay to maintain a norm of fairness (Fehr and Schmidt, 1999; Gintis et al., 2005). If this were the case, we might expect costly punishers should act fairly in other contexts, but evidence suggests costly punishment is not always correlated to prosocial behaviour (Yamagishi et al., 2012; nas Garza et al., 2014). Experimental data suggests that punishment is expressed for multiple reasons,

	Defect	Trust
Investor	1	rb
Trustee	0	$(1-r)b$

Table 5.1: The Trust Game. An investor begins with one unit of fitness. That investor then chooses to trust and invest in a trustee, or to defect. If the investor defects, then they keep their pay-off of 1, and the trustee receives 0. If the investor trusts, then the investment is multiplied by some factor b, and the trustee returns some fraction r to the investor. The investor will earn a profit if the trustee returns $r \geq 1/b$.

including both revenge and a predisposition for fairness (Bone and Raihani, 2015). Even where it benefits society, punishment is often motivated by anger (Hopfensitz and Reuben, 2009), and is known to be linked to social dominance (Pfattheicher et al., 2014; Pfattheicher and Schindler, 2015). Evolutionary game theory has shown that costly punishment can be adaptive if selective pressures focus on relative rather than absolute payoffs (Barclay and Stoller, 2014; Huck and Oechssler, 1999), or if selective pressure is weak (Rand et al., 2013).

Adding the perspective of costly punishment to the Trust Game may only seem to confuse matters further. Why would it be beneficial to reject advantageous offers on the one hand (costly punish), but accept offers of unknown — and thus potentially deleterious — quality on the other (blindly trust)? If humans are predisposed to fairness, then why is costly punishment not correlated with trusting others (Yamagishi et al., 2012)? If humans are attempting to out-compete each other through relative payoffs (Huck and Oechssler, 1999), then why act trustingly at all (Berg et al., 1995)?

Partner choice and its resulting market dynamics may solve part of this quandary (Noë and Hammerstein, 1994; dos Santos, 2014). When people can form reputations and select partners, higher levels of cooperation are witnessed in the laboratory (Sylwester and Roberts, 2013). Recent research indicates that fairness can arise from partner choice (Debove et al., 2015; André and Baumard, 2011; Sylwester and Roberts, 2010; Chiang, 2010). However, these results still do not explain why an individual would reject all unfair offers (i.e. costly punish).

Here, we demonstrate that costly punishment can best be explained in concert with trust. By assuming only a diverse population with varying rates of return, and a condition of partial information, the Trust Game is transformed into a situation of partner choice. By creating competition between (i) trustees whose individual return rates are known, and (ii) trustees with unknown individual rates, we find that blindly trusting strangers and costly punishing can both be adaptive. Further, we show that costly punishment cannot evolve in this context without trust, and such punishment is only adaptive in environments of partial information. Rejecting all known offers can be advantageous if and only if it includes the implicit threat to trust unknown offers.

5.4 The Evolution of Trust

5.4.1 Model

Manapat et al. (2012) argue that the trust demonstrated in the laboratory may be a consequence of humans evolving in small groups with partial information. Here we replicate their results using simulated evolution. We show that trust is adaptive given both partner choice and occasional knowledge of trustees' return rates.

A population consists of $N_i = 500$ investors and $N_t = 500$ trustees. A trustee is genetically encoded with $r \in [0...1]$, its *return rate*. This is the fraction of the investment the trustee will return to the investor. Each investor possesses a *trust* attribute $t \in [0...1]$, which represents the chance of the investor trusting a trustee when they possess no information regarding the trustee's return rate. For all simulations shown here, we set $b = 3$. With probability t, an investor trusts a trustee and the trustee is given $b = 3$. The trustee then returns rb to the investor and retains $(1 - r)b$ for itself (see Table 5.1). With probability $(1 - t)$, the investor does not trust and retains the one unit of fitness, leaving the trustee with nothing.

In a one-shot game where the investor has no information regarding the trustee's rate of return, the sub-game perfect equilibrium is not to trust. This changes when an investor (i) may select from one of multiple trustees, and (ii) might have some knowledge of a trustee's return rate. In this model, each investor is presented with k trustees and may invest with one. q is the probability of knowing the return rate, r, for a given trustee.

How should an investor choose among many potential partners when they have information about some trustees and no information about others? Here we slightly alter the decision rule presented in Manapat et al. (2012). Keeping in mind that the investor only makes money if the trustee's return rate is greater than $1/b$ (in this case $1/3$), we presume that an investor (i) selects the highest known return rate as long as $r > 1/b$. (ii) If no such return rate exists and the investor is trusting, they invest in an unknown return rate, otherwise (iii) they do not invest and keep the 1 unit of fitness.

This can be written formally. In a particular game, the investor will know the return rate of j out of the k trustees, based on the value of q. Consequently, an investor will select trustee i with return rate r_i with probability:

$$p(r_i) = \begin{cases} 1 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x = r_i > 1/b \\ 0 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x \neq r_i \\ \frac{t}{k-j} & i > j;\ \max\limits_{1 \leq x \leq j} r_x \leq 1/b \end{cases} \tag{5.1}$$

If the return rate, r_i, is the highest known rate, and the return rate is greater than $1/b$, then the investor will select it. If r_i is known, but there are other larger, known return rates, it will not be selected. Finally, if none of the known return rates are greater than $1/b$, then the investor will randomly select an unknown return rate with probability t. Thus, as an investor's trust increases, they are more likely to take a risk and

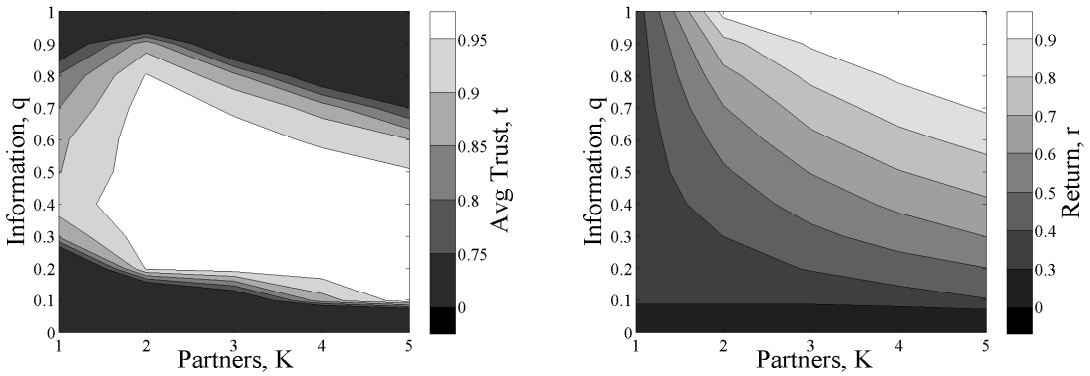


Figure 5-1: Selection for trust in the investors (left) and return rate in the trustees (right), both as functions of information the investors hold about the trustees (q) and the number of partners available in each round (k). The depicted values are averages over 10 runs with populations of 500 averaged over the final 500 generations of each run. **(Left:)** The average trust, t, in the investor population, where the investors accept the highest known return greater than $1/b$. Here $b = 3$. **(Right:)** The average return rate, r, in the trustee population.

invest in an unknown trustee.

This decision rule is employed for a couple of reasons. First, the investor is profit maximizing when information is known. If the return rates for all trustees are known ($q = 1$), the investor selects the highest return rate, presuming the highest offer is profitable ($r > 1/b$). Second, the decision rule tests the advantages of trust without the potential confound of costly punishment. Because the investor will always accept the highest known, profitable offer before risking an unknown offer, the agent never costly punishes — the investor never rejects offers that are profitable ($r > 1/b$) in order to trust. Thus, in this simulation, we test whether trust is adaptive in a profit maximizing investor who never costly punishes. In subsequent sections we will evaluate whether trust is adaptive when investors may express costly punishment.

During each round an investor is offered the opportunity to invest with one of k randomly selected trustees. Each investor plays $x = 500$ rounds of the game, and after 500 rounds, a new generation of investors and trustees are selected[1]. If an investor's trust (t) or a trustee's return rate (r), perform better compared to others, that agent and its attribute have a higher likelihood of appearing in the next generation. In line with Manapat et al. (2012), each agent is selected for the next generation using the pairwise comparison process (Traulsen et al., 2007). Initially, r and t are randomly instantiated in the range $[0...1]$ and throughout the work, $b = 3$. When an attribute is added to the next generation it is slightly mutated over a Gaussian distribution with $\mu = 0$ and $\sigma = 0.01$. $g = 1000$ generations are run, and the population's average trust and return rates are considered.

[1] All models and code for figures are (or will be on publication) available in supporting materials.

5.4.2 Results

Figure 5-1 is a replication of the findings by Manapat et al. (2012) with the novel decision rule described in Equation 5.1. Figure 5-1(left) shows that the adaptiveness of trust depends on the market size (k), and the likelihood of possessing information (q). Generally, high levels of trust evolve; investors are willing to invest with trustees despite ignorance of their individual return rate. However, a few points are worth highlighting. First, if the chance of possessing information is sufficiently low, then trusting is not advantageous. Further, trust requires more information as the market size decreases. When there is no partner choice ($k = 1$), trust begins to fail when $q < 0.3$. Finally, trust declines when the likelihood of information is high; however, this is not because trust is detrimental, only that it is not needed (see Discussion).

Figure 5-1(right) depicts the trustees' average return rate. When there is only one partner, the trustee returns a rate barely greater than $1/b$. When partner choice is added ($k > 1$), trustee return rates become a function of both k and q. When either the number of partners k or the frequency of information q increase, so does the average return rate, r.

5.4.3 Discussion

Manapat et al. (2012) showed that trust is adaptive as long as the chance of knowing at least one return rate is greater than $1/b$. Our results confirm this finding. As the number of potential partners increase, trust is adaptive despite less frequent information.

In addition, we find that as information and partner choice increase there exists a threshold where trust appears to decline (cf. Figure 5-1(right)). This is not because trust is detrimental, but rather because trust is unnecessary. When the rate of information is high, the return rate of each potential partner is likely known, thus the investor is rarely faced with the dilemma of trusting an unknown partner. Consequently, trust drifts neutrally; trusting and non-trusting investors perform similarly because they never need to trust (Manapat et al., 2012).

Figure 5-1(right) depicts the underlying market competition between the trustees. A trustee can only receive money if it is selected for investment. Without partner choice ($k = 1$), the trustee offers the minimum value which is advantageous to the investor ($r > 1/b$). As both the number of partners and the frequency of information rise, the chance increases that an investor will learn more than one return rate. As such, the trustees must raise their return rates to compete for selection. When information is fully transparent ($q = 1$), trustees are forced to offer almost everything in order to out compete other trustees.

This is line with the work by Debove et al. (2015), who show in the Ultimatum Game that partner choice can lead to large returns when there is an imbalance in the number of investors and trustees. Again, however, while trustee partner choice may generate higher return rates, it does not explain the human propensity to reject all profitable offers. We address this in the next two experiments.

5.5 Costly Punishment Evolves Due to Partial Information

Here we extend the model and demonstrate that costly punishment evolves with partner choice and partial information. We show that complete transparency of information does not lead to rejecting unfair offers. Rather, a demand for fairness evolves only when information is partially obfuscated.

In Experiment 1, an investor selected the largest known return rate, provided the rate was larger than $1/b$. As a consequence, investors were not permitted to costly punish (i.e. reject a profitable offer). We can relate the above decision rule to the notion of minimal acceptable offer (MAO), found in the Ultimatum Game. As the name suggests, the MAO is the minimal offer an investor will consider.

In Experiment 1, investors' MAO was set to $1/b$. Here, we allow the MAO to evolve. By doing this, we permit costly punishment. If an investor rejects offers greater than $1/b$, then the agent is rejecting profitable offers, which by definition is costly punishing. We show that a MAO closer to fairness $(1/2)$ is adaptive given partner choice and partial information.

5.5.1 Model

We add a new variable to each investor, demand $d \in [0...1]$. Demand is the MAO that the investor will accept when a trustee's return rate is known. An investor is now characterised by both trust, t, and its minimum acceptable offer, d. If an investor's demand is 0.5, then it will only accept offers that are fair or better — the trustee must offer at least a 50% return. If the investor's demand is $1/b$, then it behaves exactly as before.

We can formalize the investor's decision rule. Based on the transparency of information, q, an investor knows the return rate for j out of k trustees. It invests with trustee i who has return rate r_i with probability:

$$p(r_i) = \begin{cases} 1 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x = r_i \geq d \\ 0 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x \neq r_i \\ \frac{t}{k-j} & i > j;\ \max\limits_{1 \leq x \leq j} r_x < d \end{cases} \tag{5.2}$$

There are only two changes from Equation 5.1. First, a trustee with a known return rate is only chosen if its return is greater than or equal to d, rather than $1/b$. Second, based on its trust, t, an investor invests with an unknown trustee if none of the known return rates meet the MAO of the agent.

Again, this decision rule is useful because an investor is always revenue maximizing if all return rates are known. Since simulated evolution selects for the agents with the highest payouts, the decision rule creates pressure to find the revenue maximizing values of d, t, and r. This allows us to test whether costly punishment and trust can be adaptive in the limit case — when agents are attempting to maximize profit.

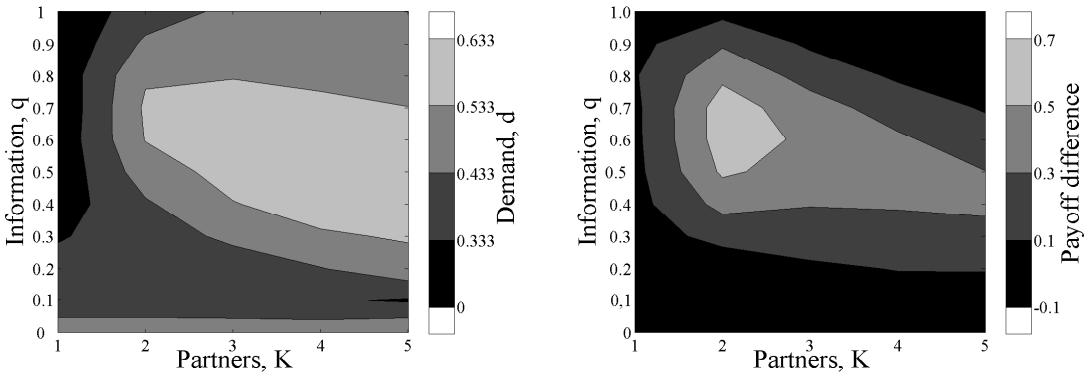


Figure 5-2: Standards of fairness and costly punishment benefit investors and are selected for in conditions of partial information and partner choice. **(left)** The investors' average minimum acceptable offer (d) over the final 500 generations, averaged over 5 runs. **(right)** Investors' average pay-off difference when demand (d) evolves versus when it is held static at $1/b$, as in the first experiment.

5.5.2 Results

Figure 5-2(left) depicts the population's average MAO (d) for differing values of q and k. Without partner choice ($k = 1$) and as long as information is moderately translucent ($q \geq 0.3$), it is not advantageous to demand a return rate greater than $1/b$. However, as soon as an investor can select between multiple partners, $k > 1$, then the optimal minimum acceptable offer increases. On average, investors are willing to reject profitable offers.

Figure 5-2(right) represents the payoff difference between investors who evolved demand and those in the previous experiment where the MAO was held static at $1/b$. Positive values represent the contexts where an investor earns more by rejecting profitable offers. Thus, in Figure 5-2(right), when the payoff difference is positive, then the ability to evolve costly punishment (i.e. MAO $> 1/3$) leads to an increased payoff for the investor, and consequently an evolutionary advantage. As the figure illustrates, it is only beneficial to reject profitable offers given partner choice ($k \geq 2$) and when information is not fully transparent ($q < 1$). Interestingly, these results, while reliable, derive from unstable dynamics, for an example run see Figure 5-3.

Finally, despite permitting the evolution of a MAO, trust still evolves. The graph is not shown due to space constraints, but the average levels of trust are similar to those depicted in Figure 5-1(left).

5.5.3 Discussion

It seems paradoxical that individuals trust unknown partners (Berg et al., 1995), but reject profitable offers when information is known (Manapat et al., 2012). Here, we demonstrate that contexts exist where such behaviour is adaptive. In small groups with partial information, there is selective pressure to both reject unfair offers and trust unknown offers.

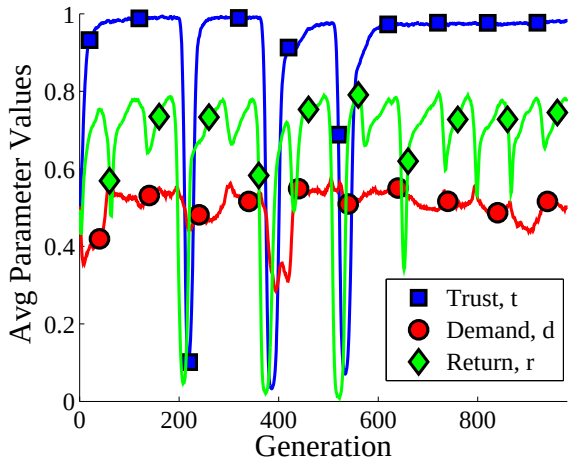


Figure 5-3: Coevolving average return rates (diamonds), trust (squares), and MAO (circles) over 1000 generations in a single exemplar run, where $k = 3$; $q = 0.5$. Trust and return rates are unstable, but reliably high on average; demand is similarly unstable but hovering near the fair value of 0.5. Defecting trustees occasionally benefit, before being culled.

Whilst rejecting profitable offers evolved throughout most of the parameter space (see Figure 5-2(left)), such behaviour did not always increase investor payoff. For instance, when the number of partners and information prevalence are high (upper right corner of Figure 5-2(right)), rejecting unfair offers is not advantageous. This is because with widespread knowledge of return rates (high information), competition between trustees pushes these rates to high levels. A willingness to reject offers of 0.5 is irrelevant if trustees are always offering returns above 0.9 (see Figure 5-1(right)). Thus, demand, d, has negligible effect and drifts neutrally.

Interestingly, there is only one context where rejecting profitable offers is advantageous — when there is both partially-occluded information ($q < 1$), and partner choice ($k > 1$). What is unique about partial information? When information is not fully transparent, there is an opportunity for an investor to trust an unknown trustee. When information is transparent, trust is never a factor because, by definition, trust is only applicable when there is risk.

Our results indicate that costly punishment is adaptive because of its impact on the marketplace. The benefit of punishment derives from its impact on the selective landscape of the trustees. Rejecting all known offers is adaptive because it provides selective pressure on the trustees, forcing the trustees to increase their offers. Why is that? To further examine this phenomenon, we next explore demand rates when trust is removed from the investors.

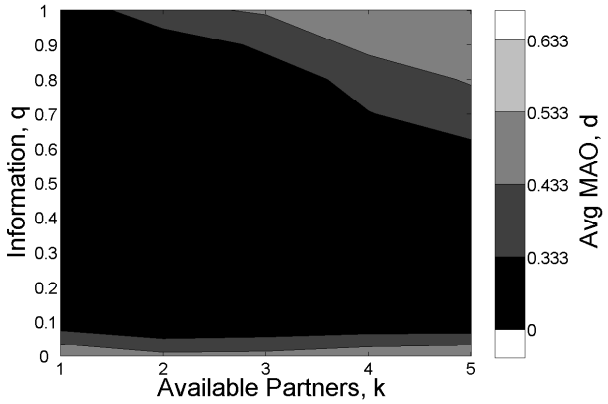


Figure 5-4: Costly punishment cannot evolve without trust. Average investor MAO (d) is shown when trust is held static at 0. Results averaged over the final 500 generations and over 10 runs.

5.6 Costly Punishment is not Adaptive without Trust

Here we analyse the interdependence between trust and costly punishment. Above, we demonstrated that both trusting and costly punishment are adaptive with partner choice under partial information. Here we show that costly punishment is not adaptive without trust.

In the previous experiments, if no trustee offered a return rate above an investor's MAO, the investor had to decide whether to trust an unknown trustee, or keep its 1 unit of fitness. Here we force all investors' trust to zero. The probability of selecting a trustee i with return rate r_i now becomes:

$$p(r_i) = \begin{cases} 1 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x = r_i \geq d \\ 0 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x \neq r_i \\ 0 & i > j \end{cases} \tag{5.3}$$

When a trustee's return rate is unknown ($i > j$), it will never be selected. If no known trustee meets an investor's MAO, the investor will simply keep its one unit of fitness.

5.6.1 Results

Figure 5-4 shows the population's average MAO, d, when investors are untrusting, $t = 0$. Generally, costly punishment does not evolve; the MAO rarely raises above $1/b$. Only when both information and the number of investors are high is $d > 1/3$.

5.6.2 Discussion

Generally, costly punishment cannot evolve without trust. However, in the upper right corner of Figure 5-4, higher minimum acceptable offers occur. This is, again, due to neutral drift in contexts where prevalent information leads to increased competition. To validate that none of the evolved demand rates conferred advantages to the investor, we ran another simulation. As before, we held trust at zero ($t = 0$), but this time we also held the demand rate static at $d = 1/3$. In such a context, the investor will never leave a profitable offer on the table, but will also never trust an unknown offer. We subtracted the average payoff from investors who evolved d from those where $d = 1/3$. No evolved investor payoff outperformed investors where $d = 1/3$ (graph not shown, because all numbers were less than zero). Consequently, even when the average MAO drifts above $1/3$ in Figure 5-4, it confers no benefit.

The finding that trust mediates adaptive costly punishment is congruent with the findings of Balliet and Van Lange (2013), who show in a global meta-analysis that punishment only promotes cooperation where levels of trust are high. However, there is a potential confound to their analysis — trust and cooperation are both correlated with wealth. It may be that trust, cooperation, and other public goods are a luxury more prevalent in societies well under the carrying capacity of their environment (Sylwester et al., 2013b). Nevertheless, the dynamics of our results hold across even the economically neutral contexts of abstract simulations, so must be considered a parsimonious explanation for this observed regularity.

In summary, rejecting profitable offers cannot evolve without trust. But, why is that? Should not high demand rates threaten the trustees? If trustees do not acquiesce to high demands, then no one will invest. However, we have shown that, in an evolutionary context, this threat is not sufficient to raise trustee returns. If simply raising demands was sufficient to increase returns, then high minimum acceptable offers would have evolved without trust.

What does trust confer which enables the rejection of unfair offers? High trust increases the likelihood that an investor will invest with a trustee despite ignorance of the trustee's return rate. If removing trust eliminates the advantages of rejecting profitable offers, then we know that it is not just the threat of the MAO which increases the return rates of the trustees. Instead, the combination of a high MAO mixed with the threat of investing with an unknown trustee increases the trustee return rates. By eliminating competition between known and unknown return rates, we have eliminated positive selection for d. Costly punishment is only adaptive in the context of co-evolving levels of demand, trust, and rates of return.

5.7 General Discussion

We have proposed a novel explanation as to why humans both trust unknown (and thus potentially disadvantageous) offers, yet reject profitable but unfair offers. In small markets with partial information, these seemingly paradoxical behaviours are adaptive. This is because creating competition between trustees with known and unknown rates of return is advantageous for the investor.

The willingness to trust an unknown partner, occurring with the willingness to reject profitable but unfair known offers generates just such a competition. Because investors are willing to risk unknown partnerships, they are able to evolve higher minimum acceptable offers. Once the competition between the unknown and known trustees is created, the trustees are forced to raise their return rates. Neither cooperation nor defection on the part of the trustees is stable, but overall cooperation is sufficiently frequent to generally provide high expectations for fairness. Even where these collapse, the dynamics of the market are such that the system rapidly recovers (Figure 5-3). However, where investors do not trust unknown offers, raising minimum acceptable offers confers no benefit. Trust is a prerequisite for the evolution of rejecting unfair offers, and, as suggested by Queller and Strassmann (2013), requires a measure of ignorance.

In our system, we see trust and costly punishment decay either when there is a high probability of knowing any one partner's return rate, or when there are simply a large number of partners providing a sufficient market of known return rates. In situations of high information, the investors—those able to multiply the investment—are able to hold on to nearly all of their profits, substantially reducing the benefit of cooperation to those who invested with them. In an age of information, this may be a matter of social concern.

Our work stands in contrast to the majority of analysis regarding the effects of partner choice on fairness, which presume the return rates of all partners are known (Baumard et al., 2013, for a review). Further, our work extends the recent discussion on how outside options affect fairness norms (Debove et al., 2015). Evolving trust increases the number of potential investments, and the value of a weak offer decreases when the investor can seek another, albeit unknown, offer. Our findings may account for at least some within- and between-population variation in what is viewed as fair (Henrich et al., 2001) and in cooperative investment more generally (English et al., 2015). Our results are also compatible with empirical findings of variation in punishment due to self-reported dominance (Pfattheicher et al., 2014), and hormone levels (Pfattheicher and Keller, 2014), as these may reflect either intrinsic or experientially-derived variation in confidence and expectation. Our model as it stands though is probably not sufficient to account for the qualitative differences in strategy observed between those who punish indiscriminately from those who punish purely to the benefit of the group (nas Garza et al., 2014; Sylwester et al., 2013b). Because the costly punishment takes the form of refusing to interact rather than active injury, our model also may evade the reputational costs of punishment reviewed by Raihani and Bshary (2015).

A potential criticism is that our results are considered over an evolutionary time-frame. Humans are not genetically, unconditionally trustworthy, we frequently adjust our strategies based on prior experience (Bear and Rand, 2016; Rand et al., 2012). Individuals are calculatingly trusting (Williamson, 1993). We agree, and suggest that the present results can be considered over an individual life-time, demonstrating the conditional nature of trust. Evolutionary algorithms are particularly useful in uncovering advantageous strategies in populations where the frequency distribution of strategies affect the outcome of each action (Alexander, 2009). They are learning algorithms, and can be metaphorically applied to learning both within and across lifetimes (McNamara and Weissing, 2010). If the evolutionary metaphor is ripped away, the algorithm still

searches for the best strategy at any given moment in time. If each "generation" is considered as an attempt to find one of the best actions given the current state of the market, then Figure 5-3 demonstrates the conditional nature of trust. Generally trust is advantageous, however, if trustees attempt to exploit investor trust, trust quickly dissipates. Importantly, trust just as quickly reappears when trustees begin to offer acceptable return rates. This matches the evidence that individuals will quickly forgive harmful transactions when it is in their benefit to do so (Fudenberg et al., 2012).

The present work illustrates the development of a parsimonious, self-sustaining institution for costly punishment, requiring only trust and partial information. Given the simplicity of this institution, our work begs the question of why there should be variation in cooperation globally (e.g. Henrich et al., 2001; Herrmann et al., 2008b; Raihani et al., 2013). A possible explanation is that the capacity for trustworthiness may be limited by local economic factors where there is high competition (Sylwester et al., 2013b; Raihani et al., 2013). Returning to trust as an institution, note that there is no intrinsic reason to expect investors and trustees to be members of different populations; they are described as such here merely for clarity. Our model requires only that individuals behave as described above when they take on these two different roles. The implied context-driven behavioural inconsistency would not be surprising in light of known decision theory (Tversky and Kahneman, 1981). Indeed, Espín et al. (2015) report an anticorrelation between return rates and demanded offers by individuals in the context of the ultimatum game, indicating that while both may be determined by personality characteristics, there is no explicit or logical congruence between behaviour in the two different roles.

5.8 Conclusion

We have demonstrated that both trusting and rejecting profitable offers is revenue maximizing, provided that minimal acceptable offers and trust coevolve. In Experiment 1, we replicated the result that blind trust is adaptive without costly punishment. In Experiment 2, we demonstrated that costly punishment (through the rejection of profitable offers) is advantageous presuming trust exists in environments of partner choice and partial information. Finally, we showed that costly punishment cannot evolve without trust. This provides a relatively simple explanation for the apparently maladaptive trait of costly punishment. Costly punishment cannot evolve without both ignorance and trust. Demanding a fair offer is only reliably advantageous if it includes the threat to accept unknown offers.

6

Heterogeneity in Costly Punishment Strategies is Adaptive

> *Knowledge kills action; action requires the veils of illusion.*
>
> Nietzsche, *The Birth of Tragedy*

6.1 Summary

This brief chapter extends the previous chapter. Whilst Chapter 5 analyses how costly punishment of unfair offers can be adaptive, it does not seek to explain the variety of strategies found in experimental data. Here we demonstrate that mixed strategies of costly punishment are adaptive given partner choice and partial information. To do this, we extend the work on the Trust Game. We find that trustees are forced to raise their return rate if investors play mixed strategies — this is true even when half the investors are willing to accept profit losing offers. When information about the best strategy is unknown or hard to acquire (i.e. one lacks veridical knowledge), varied strategies can perform almost as well as the optimal strategy.

6.2 Introduction

In the previous chapter, we demonstrated that, when playing the Trust Game (TG), a high minimum acceptable offer (MAO) is revenue-maximizing under the plausible conditions of partial information and partner choice. However, while research has begun to explain the evolutionary utility of costly punishment (Barclay

and Stoller, 2014; Huck and Oechssler, 1999; Rand et al., 2013; Debove et al., 2015), not everyone punishes with the same frequency. In fact, there is often significant variance in the minimum offers individuals will accept both within a culture and cross culturally (Henrich et al., 2005). If the costly punishment of unfair offers is adaptive, why do so many accept less than fair offers? Selective pressure for costly punishment implies that abstaining from costly punishment reduces profit, limiting the viability of the strategy. How can such variance in strategies evolve?

Manapat et al. (2012) considered the variation of minimal acceptable offers from an evolutionary perspective. They asked participants their MAO for a Trust Game where the subgame perfect strategy is for the investor to accept any offer greater than $1/3$ (see Chapter 5). Figure 6-1 represents their experimental results. Whilst a large subset of individuals possessed a fair MAO (i.e. 1/2), responses were quite varied. Manapat et al. (2012) wondered whether such variation provides a benefit to a population of investors. To test this, they showed that the best fit model for the data required individuals to possess 'fuzzy' minds, causing investors to occasionally accept and reject offers that are deleterious. However, using evolutionary game theory, they proved that fuzzy minds are not evolutionarily viable; a population of fuzzy-minded individuals are invaded by rational, profit seeking individuals. Consequently, they left the evolution of the heterogeneity of costly punishment strategies as an open question.

Whilst significant research has considered the adaptive nature of costly punishment, few have attempted to diagnose the causes underlying the heterogeneity of individual punishment strategies (Fischbacher et al., 2013). Here we demonstrate how variation in costly punishment strategies can evolve given partial information and partner choice. Extending the model in Chapter 5, we show that mixed costly punishment strategies are advantageous given partial information and partner choice. This holds even when half of the population is willing to accept unprofitable offers.

6.3 Model

We employ the same agent-based model of investors and trustees as the previous chapter. If you recall, investors are comprised of two traits: trust (t), and a MAO (d). The minimal acceptable offer represents the lowest offer the agent will accept. Trust represents the likelihood the investor will invest with an unknown trustee if no known offers meet the MAO. Trustees are comprised of one trait, their return rate (r). This is the fraction of the investment the trustee will return to the investor.

In Experiment 2 of Chapter 5, all three variables were allowed to co-evolve. We demonstrated that investors who are allowed to costly punish ($d > 1/3$) outperform those who accept any profitable offer ($d = 1/3$; see Experiment 1 of Chapter 5). We concluded that, given partial information and partner choice, it is adaptive to costly punish (i.e. $d > 1/3$).

In this experiment, rather than evolving a MAO (d), we evolve variance in an investor's MAO. As in Experiment 1 of the previous chapter, we hold d static at $1/3$. We then add a new variable, σ, which defines the variance in the investor's MAO. Each game, an investor will uniformly select its MAO in the range

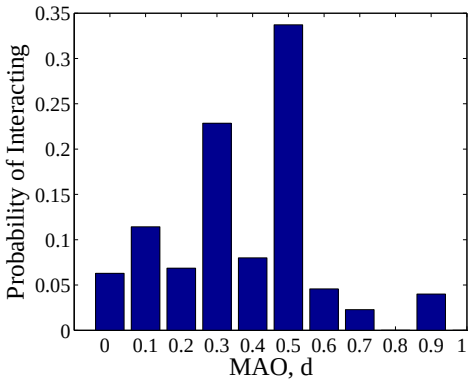


Figure 6-1: Probability density function of participants' minimal acceptable offer from Manapat et al. (2012).

$[d - \sigma, d + \sigma]$. Thus, if $\sigma = 0.1$, then during each round of play, the investor will use a different MAO, uniformly selected in the range $[0.23, 0.43]$. Consequently, sometimes the investor will costly punish (MAO $> 1/3$), but other times the agent is willing to accept unprofitable offers (MAO $< 1/3$). By permitting σ to evolve, we can test whether a mixed strategy of costly punishment is advantageous (i.e. $sigma > 0$).

As previously, during each game, an investor is able to invest with one of k trustees. There is q probability that the investor knows the return rate, r, for any given trustee. Since an investor's MAO is now defined over a range, y is the investor's current MAO selected from the range $[d - \sigma, d + \sigma]$. Since an investor knows the return rate of j out of k trustees, the likelihood that the agent will invest with trustee i with return rate r_i is defined as:

$$p(r_i) = \begin{cases} 1 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x = r_i \geq y \\ 0 & i \leq j;\ \max\limits_{1 \leq x \leq j} r_x \neq r_i \\ \frac{t}{k-j} & i > j;\ \max\limits_{1 \leq x \leq j} r_x < y \end{cases} \tag{6.1}$$

If the return rate of trustee i is both the largest known return rate and is greater than y, then the investor will accept the offer. If r_i is known, but another known return rate is greater than r_i, it will never get selected. If r_i is unknown, and no known return rate is greater than y, then r_i is selected based on the trust (t) of the investor.

6.4 Results

Figure 6-2(a) depicts the evolved values of σ over a parameter sweep of potential partners (k) and information transparency (q). When an investor can select from more than one partner ($k > 1$), σ is positive. However, positive values of σ do not always confer an advantage. Figure 6-2(b) illustrates the trustee return

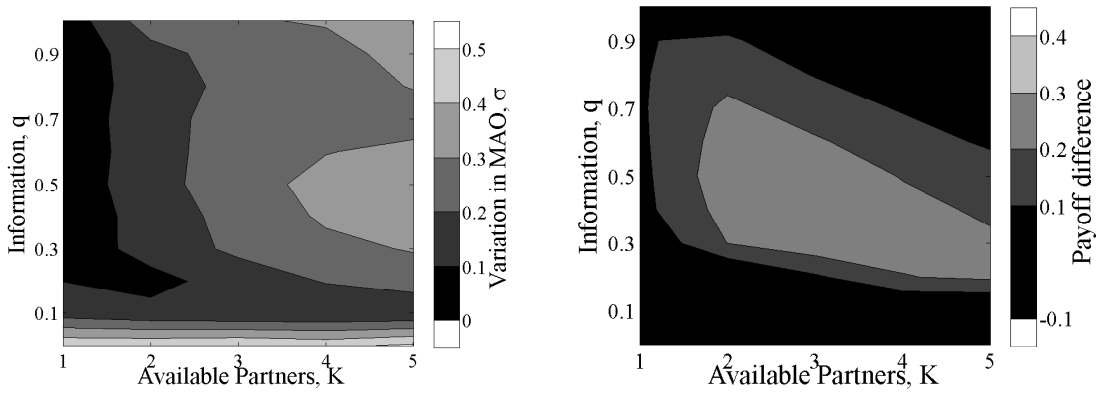


Figure 6-2: Variance in costly punishment benefits investors in conditions of partial information and partner choice. **(a)** The investors' average variance in costly punishment (σ) over the final 500 generations, averaged over 5 runs. **(b)** Investors' average pay-off difference when demand (σ) evolves versus when it is held static at 0 (d is held static at $1/3$)

rate difference between evolving σ compared to when $\sigma = 0; d = 1/3$ (see Chapter 5: Experiment 1). Variance in costly punishment ($\sigma > 0$) is profit enhancing given partner choice ($k > 1$) and partial information ($q < 1$).

6.5 Discussion

In Chapter 5 we showed that costly punishment ($d > 1/3$) is adaptive given partial information and partner choice. However this does not explain why such heterogeneity exists in individual rates of costly punishment. Here we demonstrated that variation in costly punishment strategies can be adaptive.

Rather than evolving a MAO offer, the investors evolved variance in their MAO. Every round an investor chose a MAO in the range $[d - \sigma, d + \sigma]$. Since d is kept static at $1/3$, an agent with a non-zero value of σ would accept detrimental offers ($y < 1/3$) just as often as they would costly punish ($y > 1/3$). Interestingly, this is still advantageous for the investor. Trustee return rates are forced to rise to compensate for the occasional costly punisher, despite the fact that, at any given time, half the investors are willing to accept profit diminishing agreements.

6.5.1 Lacking Veridical Knowledge

In Chapter 5 investors could learn the optimal MAO so as to maximize their revenue. In this sense, the investors used veridical information to exploit the Trust Game. Those who knew the best strategy were rewarded. In this chapter, however, agents never learned the optimal strategy. Instead, the investors employed

a variety of strategies. This had similar effects on the trustees; return rates were raised. If the optimal strategy is difficult or time-consuming to acquire, playing a mixed strategy where half the population is willing to receive detrimental offers is still better than a world without costly punishment.

6.5.2 Why Evolve σ over d

We compared trustee return rates when σ evolves versus when d evolves (see Chapter 5: Experiment 2 — graph not shown). Trustee return rates are higher when the MAO is permitted to directly evolve. Evolving variance in the MAO ($\sigma > 0; d = 1/3$) is better than not permitting costly punishment ($\sigma = 0; d = 1/3$), but is worse than evolving the MAO ($\sigma = 0; d > 1/3$). If this is the case, then why is variance in costly punishment interesting?

The adaptive nature of mixed costly punishment strategies is important for two reasons. First, it matches the experimental data. As previously noted, significant variance is found in individual rates of costly punishment (Henrich and Gil-White, 2001; Manapat et al., 2012). Historically, models have focused on explaining costly punishment of unfair offers (Barclay and Stoller, 2014; Huck and Oechssler, 1999; Rand et al., 2013; Debove et al., 2015). As a consequence, explanations for the heterogeneity of punishment strategies have been neglected (Fischbacher et al., 2013).

This work focuses on solving this conundrum. Previous work discovered that experimental data is best explained if individuals are 'fuzzy-minded' and occasionally accepted detrimental deals (Manapat et al., 2012). However, they were unable to explain how such practice could survive natural selection. Here, given partial information and partner choice, 'fuzzy-mind' strategies are adaptive. Even if half of all MAO are unprofitable, occasional costly punishment forces trustee rates higher.

The second reason this study is important is because evolving variance in MAO may side-step a well-known difficulty in evolving costly punishment — the second-order free rider problem. Whilst costly punishing removes defectors, cooperative players who refuse to punish outperform punishers because they never pay the cost to punish (Sasaki et al., 2015; Boyd and Richerson, 1992). As a consequence, those who costly punish are invaded by cooperative, but non-punishing strategies (Sigmund, 2007).

Recently, varied, stochastic behaviour has been shown to help solve the second-order free riding problem (Krasnow et al., 2015; Chen et al., 2014). Krasnow et al. (2015) demonstrated that probabilistic punishment helps avoid the problems of the second-order free riding problem. The present work supports the findings of Krasnow et al. (2015), variation in costly punishment strategies is adaptive. Considered in the light of the second-order free riding problem, the present work may go some way into explaining the evolutionary utility in the heterogeneity of costly punishment strategies found in experimental data.

6.6 Conclusion

Whilst several researchers have shown that costly punishment can be adaptive, few have explained the heterogeneity of individual strategies. We demonstrated that mixed strategies of costly punishment are adaptive in environments of partner choice and partial information. Even though half the population is willing to accept deleterious offers, occasional costly punishment is sufficient to force higher trustee return rates. This work may offer insight into how costly punishment is adaptive despite the second-order free rider problem.

7

Costly Punishment Neutrally Drifts in the Trust and Ultimatum Game

> *A peculiar feature of beliefs about politics, religion, etc. is that the private repercussions of error are virtually nonexistent, setting the private cost of irrationality at zero.*

Bryan Caplan, *Rational Ignorance vs. Rational Irrationality*

7.1 Summary

In the Ultimatum Game, individuals frequently costly punish unfair offers. Over the past decade, different theories have explained how this behaviour can be adaptive. However, an open question remains. If punishing unfair offers is adaptive, why is there such heterogeneity in individual levels of costly punishment? Here we demonstrate how variation in costly punishment can evolve. We do this by assuming what the last two decades of research has uncovered: costly punishing unfair offers is adaptive. With this assumption, we show that a variety of strategies drift into the population. It is not because such heterogeneity is adaptive, per se, but rather because different punishment strategies neutrally drift into the population without penalty. As such, we argue that it is a mistake to attempt to map experimental variance to a specific adaptive function. No one function can explain the experimental variation in costly punishment if many different strategies return the same payoff.

7.2 Introduction

In the Ultimatum Game, a *proposer* receives a monetary windfall and is tasked with splitting the resource between him/herself and a *responder* (see Table 7.1). The responder then accepts or rejects the proposed division. If the responder accepts the offer, then the windfall is divided accordingly, otherwise both players receive nothing. In some versions of the game, the responder is tasked with defining the minimum fraction of the windfall they would accept — deemed an individual's minimum acceptable offer (MAO). Despite the fact that the sub-game perfect strategy is for the responder to accept any offer greater than zero, human subjects consistently reject profitable offers, often going so far as to demand an even split of the resource (Henrich et al., 2005).

This rejecting of profitable offers has been referred to as costly punishment, since the responder rejects a profitable offer, punishing the proposer. Given that the rejection of profitable offers negatively affects the responder, significant research has considered how costly punishment could be adaptive. It has been shown that costly punishment is adaptive if relative payoffs are more important than absolute payoffs (Barclay and Stoller, 2014; Huck and Oechssler, 1999) or if selective pressure is weak (Rand et al., 2013). Further, Debove et al. (2015) found that costly punishment evolves if rejecting an offer is more harmful to the proposer. In Chapter 5, I demonstrated that costly punishment is adaptive in environments of partner choice and partial information.

While theoretical studies have begun to explain the evolutionary utility of costly punishment of unfair offers, not everyone punishes with the same frequency. In fact, there is often significant variation in the minimum offers individuals will accept (Zhao and Smillie, 2014). In the last decade, the proximate causes of this heterogeneity have been investigated. Personality traits have been correlated to individual strategies in economic games (Zhao and Smillie, 2014; Karagonlar and Kuhlman, 2013). The neurological underpinnings of punishment have been analysed (Hiraishi et al., 2015; McDermott et al., 2009). Genetic factors have been diagnosed. A study using monozygotic and dizygotic twins playing the Ultimatum Game demonstrated that 40% of variation in individual playing styles may be explained genetically (Wallace et al., 2007). Further, priming effects, such as religion, increase costly punishment in some individuals (McKay et al., 2011a).

Despite progress in understanding the proximate causes of heterogeneity in punishment strategies, insight into the ultimate causes remains an open question (Fischbacher et al., 2013; Manapat et al., 2012). If the costly punishment of unfair offers is adaptive, why will so many accept less than fair offers? Selective pressure for costly punishment implies that abstaining from costly punishment reduces profit, limiting the viability of the strategy. How can the heterogeneity in individual strategies be maintained?

Manapat et al. (2012) considered the heterogeneity of minimal acceptable offers from an evolutionary perspective, wondering whether such variation provides a benefit to a population of responders. After asking responders their individual MAO, Manapat et al. (2012) showed that the best fit model for the data required individuals to possess 'fuzzy' minds, causing responders to occasionally accept and reject offers that are deleterious. However, using evolutionary game theory, they proved that fuzzy minds are not evolutionarily

viable; a population of fuzzy-minded individuals are invaded by rational, profit seeking individuals. Consequently, they left the evolution of the variation in costly punishment strategies as an open question.

Preliminary attempts to elucidate the underlying cause for the heterogeneity of individual strategies have focused on finding contexts where frequency dependant strategies coexist. McNamara et al. (2009b) found that, in environments where individuals can pay for information, heterogeneous strategies of prosociality coexist. However, despite preliminary studies on the underlying forces of such heterogeneity, there is a call for more in-depth study (Wolf, 2013; Fischbacher et al., 2013)

Here we demonstrate how variation in costly punishment can evolve in both the Ultimatum Game and a version of the Trust Game which is similar to the Ultimatum Game. The key is to take seriously the last two decades worth of theory demonstrating that costly punishment of unfair proposals is adaptive (Debove et al., 2015; Barclay and Stoller, 2014; Huck and Oechssler, 1999; Rand et al., 2013). Assuming this, we show that the frequency distribution of a population's MAO may vary considerably without negatively impacting a responder's payoff. If a proposer is attempting to choose an income maximizing offer (IMO) when playing against a population of responders, then once a certain number of responders demand a fair return, other players can possess lower MAOs without penalty. Given rational proposers, evolution cannot stop large variation in responder MAOs because various strategies neutrally drift without penalty.

This result is more parsimonious than previous attempts to understand the heterogeneity of individual strategies. Previous explanations attempted to map the frequency distribution of the strategies in a population to some function (Manapat et al., 2012; McNamara et al., 2009b), the idea was that the frequency distribution of a population was maintained by some force. In contrast, this work demonstrates that, presuming it is advantageous to costly punish unfair offers, significant strategic variation will drift into the population. No particular function will define the frequency distribution of strategies within the population, because a random number can drift in at a particular time.

We believe this work helps clarify the confusion in attempting to explain the variation of individual minimal acceptable offers. We argue that the heterogeneity is not adaptive; rather, individuals of many different population distributions receive identical payoffs. If there is no selection pressure between several distributions of minimal acceptable offers, then it is a mistake to attempt to map a function to the levels of variation in a game.

We conclude by analysing an underlying assumption of much of the Ultimatum Game literature — that reducing a population's mean MAO should lead to a reduced payoff for the responders. The mediating variables of a population's mean minimum acceptable offer have been extensively studied in the Ultimatum game (Schmitt et al., 2008). It has been shown that a population's mean MAO may shift for several reasons, including the amount of money at stake (Novakova and Flegr, 2013), whether responses are delayed (Grimm and Mengel, 2011), when the question is posed (Oxoby and McLeish, 2004), and cultural differences (Henrich et al., 2005).

We demonstrate the error in assuming a population's mean MAO is a proxy for the population's payoff. Two populations can receive the same payoff, despite differences in the mean MAO. We employ experimen-

	Reject	Accept
Responder	0	rb
Proposer	0	$(1-r)b$

Table 7.1: In the Ultimatum Game, a proposer begins with a monetary windfall, b. The proposer then offers some fraction, r, to a responder. The responder can then accept or reject the offer. If the responder accepts then they receive that fraction of the windfall, leaving the rest for the proposer. If the responder rejects the offer, both players receive nothing.

	Defect	Trust
Investor (Responder)	1	rb
Trustee (Proposer)	0	$(1-r)b$

Table 7.2: In the Trust Game, an investor begins with one unit of fitness. The investor then chooses to trust or defect with a trustee. If the investor defects, then it keeps its payoff of 1, and the trustee receives 0. If the investor trusts, then the investment is multiplied by some factor b, and the trustee returns some fraction r to the investor. In this version of the game, where the investor knows the return rate of the trustee, then the game morphs into a version of the Ultimatum Game (Tarnita, 2015). The investor (responder) decides whether to accept a return rate, r, or to walk away.

tal data from the Ultimatum Game to demonstrate that large variation in the distribution of a population's mean MAO does not affect the responder's expected payoff. The results of several recent papers rely on altering a population's mean minimum acceptable offer; however, we demonstrate that their results do not modify the expected payout for the participants. If individuals in two populations expect the same payout, then attempting to explain the difference between two populations is confounded. We argue that in contexts where a responder's expected payoff is stable, variation in individual minimal acceptable offers does not require explanation.

7.3 Variation in MAOs Goes Unpunished: A Toy Example

In this section we demonstrate that variation in individual MAOs may not affect the expected payoffs of responders. As a consequence, market pressure cannot select for a particular distribution of MAOs. To do this, we present a toy example of the Ultimatum Game.

We presume a society of responders exists, where each responder posseses a MAO, d. In this example, individual MAOs are distributed according to Figure 7-1(a). Half the population will accept any offer, $d = 0$, whilst the other half of the population demands at least a fair return, $d = 0.5$. During each game a rational, income-maximizing proposer is matched with a random responder. We define the probability function $P(d)$ as the likelihood a proposer will meet a responder with a minimal acceptable offer d. Given the distribution in Figure 7-1(a), $P(0.0)$ and $P(0.5) = 50\%$.

During the course of play, the proposer plays a game against each responder in the population. The proposer's average expected score per game (π_p) can then be written as a function of its return rate, r:

$$\pi_p(r) = b(1-r) * \sum_{d \leq r} P(d) \tag{7.1}$$

A responder accepts an offer when the proposer's offer, r, is greater than or equal to the responder's minimum acceptable offer, d. If a responder accepts the offer, the proposer receives the fraction of the reward which was not returned, namely $b(1-r)$. Since the proposer plays against each responder, this is multiplied by the fraction of the responders who accept the offer, $\sum_{d \leq r} P(d)$.

Depending on a responder's MAO, d, and the proposer's return rate, r, each responder receives a payoff (π_{resp}) of:

$$\pi_{resp} = \begin{cases} br & d \leq r \\ 0 & d > r \end{cases} \tag{7.2}$$

If the proposer's offer is greater than or equal to the responder's MAO, then the responder accepts the offer, receiving br. Otherwise, the proposer rejects the offer, garnering no return.

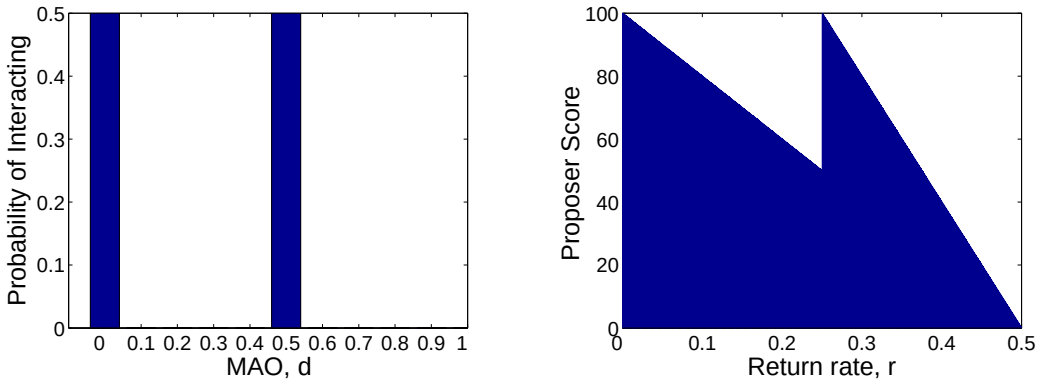
The proposer's goal is to pick the return rate which maximizes their score — the income maximizing offer (IMO). We presume that, over time, the proposer learns the frequency distribution of responder minimal acceptable offers, but the proposer does not know the MAO of the responder at a given time. Consequently, the proposer is faced with a dilemma. If a low return rate is offered, then the proposer will receive a high return from responders who accept the offer; however, few may acquiesce to such unfair offers. On the other hand, if the proposer selects a high return rate, many will accept the offer, but each trade will garner less profit.

If the proposer is profit-maximizing, they will choose a return rate, r_{imo}, which maximizes Equation 7.1:

$$r_{imo} = \arg\max_r \pi_p(r) \tag{7.3}$$

7.3.1 Results

Figure 7-1(b) depicts the proposer's expected payoff for various offers, r when playing against a population of 200 responders. If a proposer is attempting to find the optimal return rate, (r_{imo}), then two strategies perform equally: $r = 0.0$ and $r = 0.5$. Despite the fact that half the population accepts any offer (see Figure 7-1(a)), it is still just as beneficial for the proposer to offer a fair return as it is to keep everything. A proposer who keeps everything ($r = 0.0$) receives the maximum payout from the half of the population that will accept any offer. A trustee who offers a fair return ($r = 0.5$) only keeps half the maximum payout per game, but trades with everyone in the population.



(a) Distribution of MAO (d).

(b) Expected payout for r.

Figure 7-1: We consider a population where half the responders demand of MAO of $1/2$ whilst the other half are willing to accept any offer. **(a)** The probability distribution function, $P(d)$. **(b)** The proposer's expected income for a range of return rates when playing against a population of size 200.

Importantly, whichever of the two IMO strategies the proposer selects, all responders receive the same payoff. If the proposer offers the IMO strategy of $1/2$, then all participants whose MAO is less than or equal to half, will receive a return of half. An investor whose MAO is $1/2$ will accept the offer, garnering half the windfall. However, an irrational investor who will accept any offer, even zero ($d = 0$), will also accept the trustee's IMO offer of $1/2$, garnering half the windfall. If the proposer offers nothing $r = 0$, then half the responders will accept, receiving nothing, and the other half will reject the offer, receiving nothing.

7.3.2 Strategies Neutrally Drift if Enough Punish Unfair Offers

If the population's MAO distribution is altered slightly, such that 49% of the population accept any offer, and 51% demand fairness, then a fair offer is the proposer's best strategy, $r_{imo} = 0.5$. As a consequence, the fraction of the population which accepts any offer receives the same return as the rest of the population, half. In fact, if 51% of a population demand a fair offer, then it can be shown that the rest of the population can demand any value lower than half and everyone will still receive the same fair return.

Presume that 51% of the population have a fair MAO ($d = 0.5$) and 49% of the population have a MAO of x, where $x < 0.5$. Presuming $b = 1$, if a proposer offers a return of $r = 0.5$, then their average expected payoff according to Equation 7.1 is:

$$\begin{aligned} \pi_p(0.5) &= b(1-0.5) * \sum_{d \leq r} P(d) \\ \pi_p(0.5) &= 1 * 0.5 * 1 \\ \pi_p(0.5) &= 0.5 \end{aligned} \tag{7.4}$$

Since all responders are willing to accept a return rate of 0.5, the proposer garners half the windfall from each responder. Now, we check to see if any return rate outperforms $r = 0.5$.

$$\pi_p(r) \geq \pi_p(0.5) \tag{7.5}$$

Substituting from Equations 7.1 and 7.4 we get:

$$b(1-r) * \sum_{x \leq r} P(x) \geq 0.5 \tag{7.6}$$

If $r < 1/2$, then at least 51% of the population will reject the offer. However, to analyze whether there exists an r which is better than $r = 1/2$, we presume the limit case where, regardless the value of r, the other 49% of the responders automatically accept the proposer's offer. As such, the inequality simplifies to:

$$b(1-r) * .49 \geq 0.5 \tag{7.7}$$

Presuming $b = 1$, with a little simplification we find the solution:

$$r \leq -\frac{1}{49} \tag{7.8}$$

Since the return rate cannot be negative, there exists no return rate which offers a higher expected payout to the proposer. Regardless the MAO distribution of the other 49% of the population, the best offer will always be $r = 1/2$.

7.3.3 Discussion

This calls into question the validity of attempting to explain the heterogeneity in MAOs below the IMO. The last two decades of research has demonstrated that costly punishing unfair proposers is likely advantageous (Debove et al., 2015; Barclay and Stoller, 2014; Huck and Oechssler, 1999; Rand et al., 2013). If we take this assumption seriously, then a large fraction of the population should demand fair offers ($d = 1/2$). However, this does not mean that all responders must possess identical strategies. Once enough responders hold a fair MAO ($d = 1/2$) such that $r_{imo} = 1/2$, the rest of the population may hold any MAO $\leq 1/2$ without penalty. The particular distribution will not necessarily map to any function.

This proves there are contexts where heterogeneity in minimum acceptable offers cannot be selected against. If individual MAOs are distributed such that IMO = $1/2$, then any responder with $d \leq 1/2$ is expected to receive a payoff of $1/2$. Since the payoffs are identical for a wide range of strategies, evolution cannot select against any particular strategy. In our toy example, if 51% of the population have a MAO of half ($d = 0.5$), then 49% of the population may possess any return rate less than half and receive the same payoff as all other responders.

7.4 Variation in Individual MAO Goes Unpunished: An Empirical Example

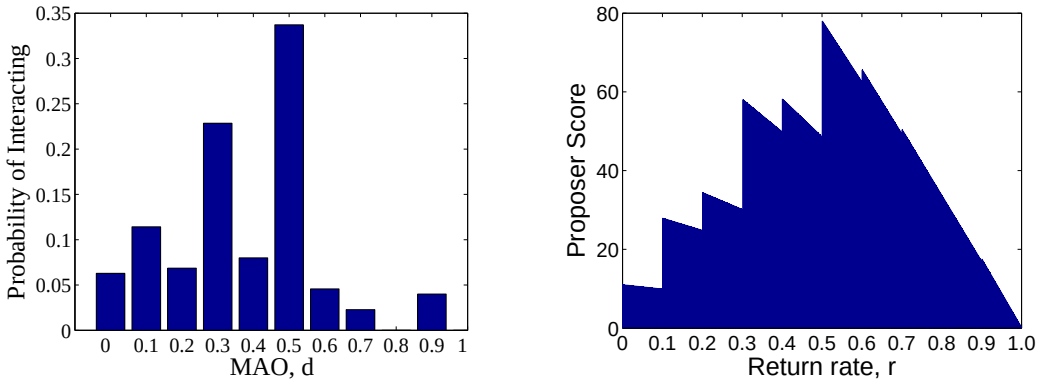
Here we consider an empirical example involving a version of the Trust Game which is similar to the Ultimatum Game (see Chapter 5). In the Trust Game, an *investor* is given a unit of money and can either invest it with a *trustee*, or leave with the money (see Table 7.2). If they invest, then the money is multiplied by some factor, b, and the trustee can decide to return some fraction of the windfall, r, to the investor. In its canonical version, the investor does not know the trustee's return rate. Thus, the investor must decide whether to trust the trustee with the investment.

Recently, research has considered the effects of offering the investor some knowledge regarding the trustee's eventual return rate (Tarnita, 2015; McNamara et al., 2009a; Manapat et al., 2012; Manapat and Rand, 2012). If the investor has knowledge of the trustee's return rate prior to deciding whether to trust, then the game morphs into a version of the Ultimatum Game (Tarnita, 2015). The investor (i.e. the responder in the Ultimatum Game) knows the return rate of the trustee (i.e. the proposer) and can either reject or accept the offer. For consistency purposes we will refer to the investor as the responder, and trustee as the proposer.

The main difference between the Ultimatum Game and this version of the Trust Game, is that the responder starts with the monetary windfall and can keep it by rejecting the offer. Thus, if the responder rejects the offer then rather than both players receiving zero, the responder receives 1 unit of money. This affects the subgame perfect strategy of the investor. In the Ultimatum Game, the subgame perfect strategy is to accept any offer greater than 0. In this version of the Trust Game, the best strategy is to accept any offer where the responder would turn a profit, namely $r > 1/b$.

Manapat et al. (2012) ran a Trust Game experiment, asking responders their MAO. In their game, if a responder invested, the windfall was multiplied by $b = 3$. Thus, the subgame perfect strategy was to accept any offer over $1/3$. Figure 7-2(a) depicts the results. Less than 25% of the population chose the subgame perfect strategy, and the most frequently selected MAO was $1/2$, a fair offer. More importantly, however, is the wide variety of responses. Minimal acceptable offers range from accepting any offer ($d = 0$) to rejecting offers less than 0.9.

Here we demonstrate how such variation can neutrally drift into the population. Once a certain number of investors demand a fair offer, other players can demand lower thresholds without penalty. Given rational



(a) Distribution of MAO (d)

(b) Expected payout for r

Figure 7-2: Using data from Manapat et al. (2012). **(a)** Probability density function of the minimal acceptable offers. **(b)** Expected proposer payout for a range of return rates.

proposers, the market cannot stop large variation in investor demands.

7.4.1 Results

As before, the best offer against the responder population in Figure 7-2(a) can be calculated. Presuming a proposer plays against each responder from Manapat et al. (2012), Figure 7-2(b) depicts the expected payouts for all possible return rates. Proposers score highest if they offer a fair return rate, $r_{imo} = 0.5$. Despite the fact that 55% of the population is willing to accept offers below an equal split, equality is the best proposal for someone playing against the entire population.

As a consequence, if the proposer employs the IMO strategy of $1/2$, then all participants with a MAO less than half also receive a return of half. The only investors who lose out are those whose MAOs are greater than half. Only approximately 10% of the population in Manapat et al. (2012) made such MAOs. Presuming a rational proposer, 90% of the population garners an identical profit of $1/2$, the other 10% receive nothing.

7.4.2 Discussion

Manapat et al. (2012) attempted an evolutionary explanation of the heterogeneity witnessed in Figure 7-2(a). Since the subgame perfect strategy for their version of the Trust Game is $1/3$, they attempted to explain why some responders accept offers less than $1/3$, whilst others reject offers greater than $1/3$. They showed that 'fuzzy-minded' responders (i.e. individuals who occasionally make suboptimal choices) provide a best-fit function for the variation in Figure 7-2(a). However, they could not demonstrate how such 'fuzzy-minded' investors could evolve.

Here we solve this conundrum in two steps. First, research has shown that humans may have evolved in

environments where minimal acceptable offers of $1/2$ are adaptive (Rand et al., 2013; Huck and Oechssler, 1999; Debove et al., 2015) — also see Chapter 5. If we presume that the best evolutionary strategy is $d = 1/2$, rather than $1/3$, then the variance which needs explaining looks quite different.

If $d = 1/3$ is the optimal strategy, then approximately 50% of the population possess MAOs both below and above the optimal strategy. As such, a 'fuzzy-minded' responder seems a reasonable explanation for the variance. However, if we presume $d = 1/2$ is an adaptive strategy (as shown in Chapter 5), then 90% of participants either chose the optimal strategy, or a MAO less than 0.5.

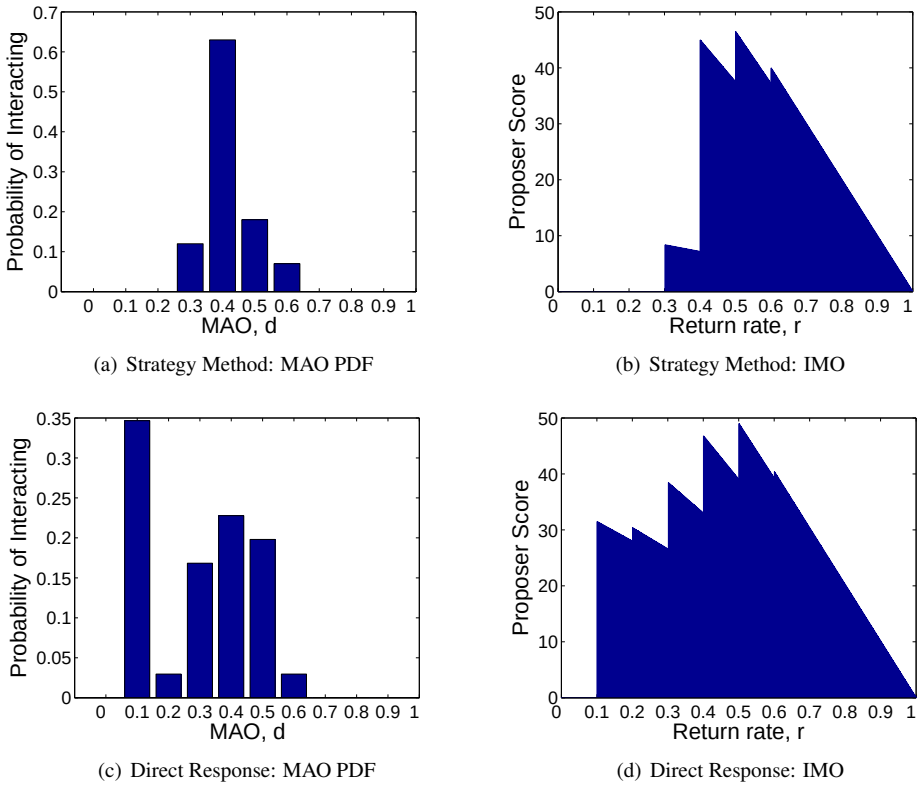
As demonstrated in Section 7.3, assuming a rational proposer, responders with an MAO less than the income maximizing offer, will receive identical payoffs compared to any other MAO less than or equal to the IMO. If 90% of the investor population is expected to receive the same payoff, then, from an evolutionary perspective, there is no selective pressure against any of the strategies with a MAO $\leq 1/2$. Presuming a fair MAO is adaptive, the market cannot stop smaller minimal acceptable offers from drifting into the population. If a sufficient percentage of the population demand a MAO $= 1/2$, such that the income maximizing offer is $1/2$, then the distribution of the rest of the population's MAO can drift lower than half.

In summary, we have shown that attempting to map a function onto the variance of individual MAOs may be confounded. A payoff maximizing proposer should not consider their offer based on one individual, but rather against the whole population of responders. If enough individuals demand fairness, then the optimal strategy is to offer fair returns. If various responders then receive identical payoffs, then they cannot be selected against. Individual MAOs can neutrally drift as long as the entire population continues to generate the same IMO.

7.5 A Population's Mean MAO is a Poor Proxy for Expected Payoff

Returning to the Ultimatum Game, significant research has sought to discover the variables which mediate an individual's minimum acceptable offer. Several of these works have then used the population's average MAO to define their results. The underlying assumption is that if, on average, a responder is willing to demand less, then they will receive less.

Here we analyse the validity of this assumption and find that a population's mean MAO is a poor proxy for its expected payoff. We consider several experimental results where the authors attempt to explain changes to the distribution of a population's MAO. Though the mean MAO shifts in different experimental settings, we demonstrate that the best strategy of the proposer (i.e. the IMO) remains unchanged, despite experimental manipulation. Consequently, responders with a MAO less than the IMO should go unpunished, enabling such strategies to drift into the population. We conclude with a discussion on the error in presuming proximate mechanisms for shifts in MAOs imply that the behaviour underwent evolutionary selective pressure.



(a) Strategy Method: MAO PDF

(b) Strategy Method: IMO

(c) Direct Response: MAO PDF

(d) Direct Response: IMO

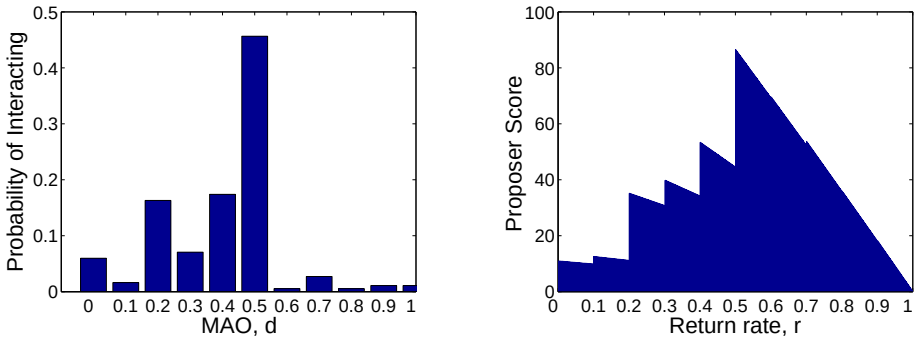
Figure 7-3: Oxoby and McLeish (2004) tested whether defining one's MAO before (strategy method) or after (direct response method) an offer affects rejection rates. The left figures represent the probability distribution functions for the two pools. The right figures depict the expected payouts for a proposer. **(a,b)** Individuals select their MAO via the strategy method. **(c,d)** Individuals select their MAO after via the direct response method.

7.5.1 Strategy v Direct-Response Method

A responder's minimum acceptable offer can be acquired in two ways. With the strategy method, proposers are asked their MAO prior to receiving any offers (Mitzkewitz and Nagel, 1993). With the direct response method, a responder's MAO is calculated via the responder's play over time. If the responder ends up rejecting offers less than some value, d and accepting offers over d, then d represents the responder's MAO.

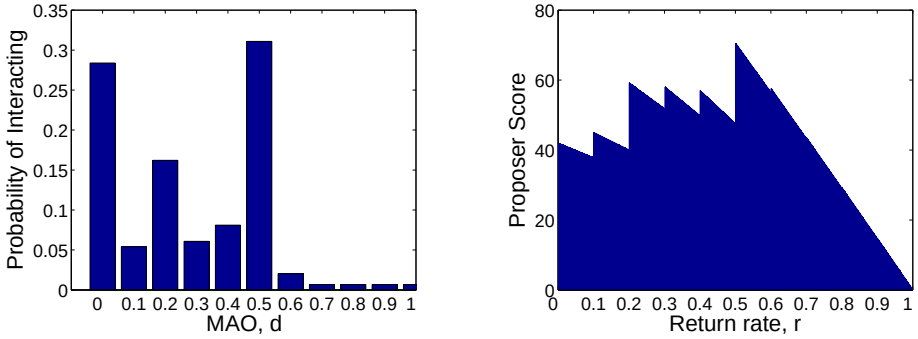
Oxoby and McLeish (2004) analysed the effect of the two methods on minimum acceptable offers (d). Figures 7-3(a) and 7-3(c) illustrate that the different methods cause fairly disparate MAO distributions. When defining their MAO before the offer, Figure 7-3(a), the majority of participants request a return rate of at least 0.4. The population's mean MAO is $\bar{d} = .42$. In contrast, with the direct response method, individuals accept much lower offers, Figure 7-3(c). The population's mean MAO is $\bar{d} = .302$.

Interestingly, however, while the distribution of MAOs change, the proposer's optimal offer (IMO) is



(a) Low Stakes: MAO PDF

(b) Low Stakes: IMO



(c) High stakes: MAO PDF

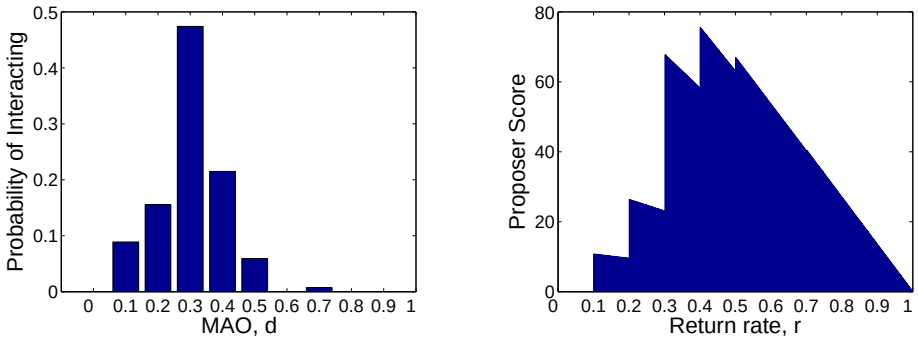
(d) High Stakes: IMO

Figure 7-4: Novakova and Flegr (2013) tested the effects of high and low stakes on rejection rates. **(a,b)** Playing for low stakes. **(c,d)** Playing for high stakes.

identical when playing against either population. Figures 7-3(b) and 7-3(d) depict that fairness, $r = 0.5$ is the optimal return rate for both populations.

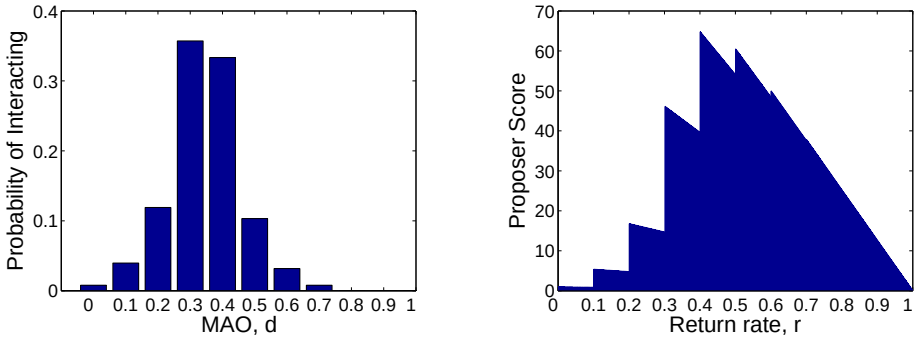
7.5.2 Monetary Stakes

Novakova and Flegr (2013) hypothesized that individuals will accept lower offers if the monetary stakes are increased. Figures 7-4(a) and 7-4(c) lend credence to this proposition. When playing for small stakes, the mean MAO is $\bar{d} = .401$, whilst $\bar{d} = .279$ when playing for larger stakes. However, again, the different distributions are not enough to alter the IMO. Figures 7-4(b) and 7-4(d) illustrate that the proposer's optimal offer remains $r_{imo} = 0.5$.



(a) Global Competition: MAO PDF

(b) Global Competition: IMO



(c) Local Competition: MAO PDF

(d) Local Competition: IMO

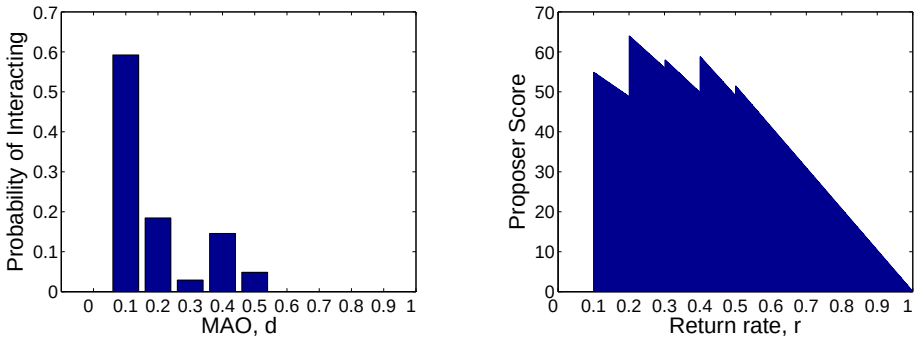
Figure 7-5: Barclay and Stoller (2014) tested the effects of global and local competition on rejection rates. **(a,b)** Absolute payoffs are highly valued. **(c,d)** Relative payoffs (i.e. local competition) are rewarded.

7.5.3 Local v. Global Competition

It has been argued that costly punishment is adaptive when there is selective pressure for local competition over absolute payoffs (Huck and Oechssler, 1999). If competing with the relative payoffs of your neighbour is more advantageous than the absolute payoff of the game, then it can be beneficial to reject profitable offers in order to prevent the other player from receiving a larger payoff. Barclay and Stoller (2014) experimentally tested whether a population's mean MAO is affected by manipulating the value of local competition.

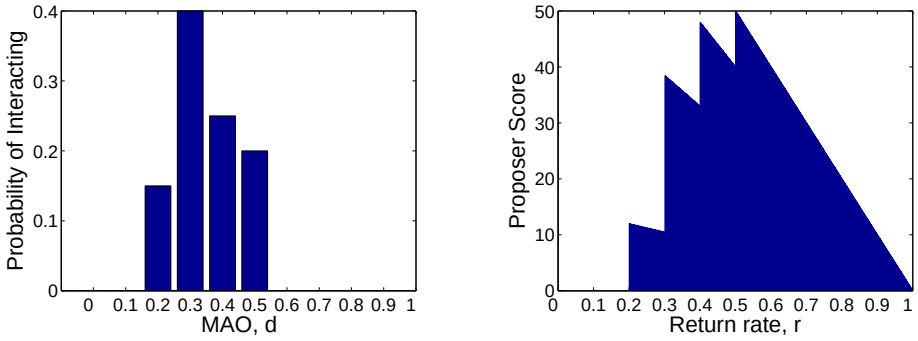
When local competition was reduced, increasing the value of absolute payoffs (i.e. global competition), then the participant's mean MAO was $\bar{d} = .3030$. When local competition was introduced, the mean MAO rose to $\bar{d} = .3444$. Barclay and Stoller (2014) concluded that there is evidence that individuals may have evolved to differentiate between environments of local and global competition.

Figures 7-5(a) and 7-5(c) depict the frequency distribution given global and local competition, respectively. Despite disparate mean MAOs, Figures 7-5(b) and 7-5(d) show that the income maximizing offer against both populations is 0.4. Presuming a rational proposer, any responder with a MAO less than or equal



(a) Delay: MAO PDF

(b) Delay: IMO



(c) Immediate: MAO PDF

(d) Immediate: IMO

Figure 7-6: The top depicts responders with a 10 minute delay, whilst the bottom Figures represent responders who answered immediately. **Figures (a,c):** the frequency distribution of minimal acceptable offers for the experimental data in Grimm and Mengel (2011). **Figures (b,d):** the IMO when playing against the population.

to 0.4, receives an identical payoff. In the global case, 85.7% of the population possessed a MAO ≤ 0.4. Under local competition 93.4% demanded a MAO ≤ 0.4.

7.5.4 Delayed Responses

Importantly, there are examples where the distribution of MAOs are disparate enough to generate changes in the IMO. Grimm and Mengel (2011) tested the effects of delaying a response to a proposal. One group of responders waited 10 minutes before accepting or rejecting an offer, whilst the other responded immediately.

Figure 7-6 illustrates their result. Calculating the IMO for both the delay, Figure 7-6(b), and immediate, Figure 7-6(d), groups, it is clear that delaying the responder's action alters the IMO. When responders are delayed, they are more likely to accept lower offers. A proposal of 0.2 maximizes the proposer's profit. In contrast, if responders respond immediately to a proposal, then the IMO is a fair offer of 0.5.

7.5.5 Discussion

Significant research has gone into explaining the distribution of a population's MAO. The implicit presumption seems to be that if an individual will accept a low offer, then that person will receive a reduced payoff. We show that this is not the case, *a priori*.

An individual exists within a population, and the distribution of that population defines the performance of the individual. While the mean MAO may shift significantly, the best strategy against that population (IMO) may remain static. This offers insight into why responders may not play the optimal strategy — they are not punished. If the best offer is fairness, $r_{imo} = 0.5$, then, presuming a rational proposer, every responder who accepts an offer less than ($d \leq 0.5$) receives the same payoff.

In the Ultimatum game, many studies have considered the variables which mediate an individual's MAO (Oxoby and McLeish, 2004; Novakova and Flegr, 2013; Grimm and Mengel, 2011; Barclay and Stoller, 2014). Many of these focus on how the population's mean MAO changes given experimental manipulation. We have shown that, despite significant fluctuations in a populations mean MAO, several of these studies do not affect the responder's expected payoff.

While these studies are valuable in understanding when and how individuals select their minimum acceptable offer (i.e. proximate mechanisms), it would be a mistake to study this behaviour from an evolutionary perspective (ultimate mechanisms). In order to discuss the selective pressures behind individual differences in MAO, a more thorough investigation of individual payoffs is required. We have demonstrated that MAO is a poor proxy for expected payoff. If proximate mechanisms are discovered which lead to reduced minimum acceptable offers, it does not, *a priori* lead to reduced individual payoffs.

7.6 General Discussion

7.6.1 Do Proposers Play the IMO?

Throughout this work we assumed rational, income maximizing proposers. We then calculated the responder's expected payoff accordingly. However, individuals are rarely rational; what if proposers offer significantly less than the IMO?

While proposers rarely play the optimal strategy (r_{imo}), rather than acting stingily and proposing offers less than the IMO, cross-cultural studies demonstrate that, if anything, proposers tend to offer more than the IMO (Lamba and Mace, 2012; Henrich et al., 2005). Further, over time, (Achtziger et al., 2015) showed that proposers raise (rather than lower) their offers because of ego depletion from fear of rejection. From a market perspective, this means that even more strategies can operate without any loss to expected payoffs.

7.6.2 If Others Demand Fairness, Why not Lower Your MAO?

What if the random proposer does offer less than the IMO? If the IMO is 0.5, but some proposers offer 0.4, it is better to have a MAO which accepts any offer. In fact, as per the subgame perfect strategy, the best MAO is still always anything above zero. As a result, the responder will never miss a potentially profitable offer. However, if everyone followed this logic the result is a tragedy of the commons. While a few people can lower their MAO in order to catch a few suboptimal offers, if everyone employs this strategy, then the optimal return rate (IMO) lowers and everyone is punished. Why is the tragedy of the commons not present in this work?

Our model is different from previous explanations of the heterogeneity of Ultimatum strategies, in that it takes seriously the idea that a high MAO is adaptive. Over the last two decades, many theories have described why this might be the case (Debove et al., 2015; Barclay and Stoller, 2014; Huck and Oechssler, 1999; Rand et al., 2013). Here we do not presume that any one theory is correct, only that a high MAO is adaptive. Without this assumption, it is true that a lower MAO would outperform a higher MAO given suboptimal or noisy play on the part of the proposer. This is exactly the social dilemma which generated such an interest in how humans play the Ultimatum Game (Sasaki et al., 2015; Boyd and Richerson, 1992; Sigmund, 2007).

However, by assuming there is some advantage for increasing one's MAO above zero, we sidestep the tragedy of the commons conundrum. Whilst there exists a pressure to lower one's MAO in the face of noisy or suboptimal proposers, we are also presuming there is some pressure to raise one's MAO. Here we do not discuss the nuances of that dynamic directly; however, as an example, consider the results from Chapter 5. In environments of partner choice and partial information, it is beneficial for responders to demand fair offers. If some portion of responders realize this, they can force return rates higher. However, not all responders must find the best MAO strategy in order to force higher return rates (see Chapter 6). The rest of the population can reap the benefits of the higher return rates, as long as their minimal acceptable offer is below the going market rate.

In this work, we focus on analysing why heterogeneity in experimental levels of MAOs persists. Presuming there is some mechanism which generates an advantage for maintaining a large MAO, then we demonstrate that strategies lower than the IMO may be able to neutrally drift into the population. This is the most parsimonious explanation of variation in UG strategies because it does not require any additional variables other than what is increasingly becoming evident, costly punishment of unfair offers is adaptive.

7.6.3 Lacking Veridical Knowledge

Throughout this dissertation, I have shown that lacking veridical knowledge can be adaptive in a variety of contexts. Here, I demonstrate an environment where incorrect beliefs are not advantageous in isolation, but are unpunished given the accurate beliefs of others in the population. First, I presume that costly punishment of unfair offers is adaptive (see Chapter 5). Thus, a MAO of 0.5 is advantageous, and consequently, possess-

ing that strategy can be considered veridical knowledge. In this Chapter, I demonstrated that if some subset of the population is unaware of the optimal strategy, their behaviour cannot be removed from the population because the best play against the whole population is still IMO $= 0.5$. Whilst lacking veridical information is not adaptive, it cannot be removed from the population as long as enough players possess veridical information. Individuals with incorrect information do not pay the cost for their suboptimal actions.

7.7 Conclusion

Here we offer a parsimonious explanation of why variation persists in individual minimal acceptable offers. We start with the assumption that the costly punishment of unfair offers is advantageous. We demonstrate that, presuming enough players are willing to punish inequity, a variety of other strategies are free to drift into the population. As a result, we suggest that the quest to attempt to uncover a variable which offers the ultimate explanation for individual differences is confounded. Further, we considered this result in the light of several empirical experiments demonstrating changes in a population's mean MAO. We suggest that changes to a population's mean MAO does not *a priori* offer insight into how and when the market should shift.

8

A Mixture of Formal and Less Formal Reviews Leads to the Best Patient Care in Inefficient Healthcare Institutions

> *One day everything will be well, that is our hope.*
> *Everything's fine today, that is our illusion*

Voltaire, *Poème sur le désastre de Lisbonne*

8.1 Summary

The benefits of whistleblowing in the healthcare sector are significant. The assumption that employee whistleblowing is always beneficial, however, is rarely examined. Here we consider how the efficacy of internal, formal inquiries in response to employees' concerns affect the utility of whistleblowing. While recent research has begun to consider how the complex nature of healthcare institutions impact whistleblowing rates; few have investigated whether institutional processes impact the benefit of whistleblowing. We find that, given resource limitations and review inefficiencies, it can actually improve patient care if either whistleblowing rates are limited or information transparency is occluded. Utilizing computer simulations, we demonstrate that a mixture of formal and less formal reviews may lead to the best patient care. We conclude with a call for further research on a more holistic understanding of the interplay between organizational structure and the benefits of whistleblowing to patient care.

8.2 Introduction

In the healthcare sector, few disagree that both informational transparency and the freedom for employees to report bad practice (i.e. whistleblowing) have promoted improved patient care. In the long term, whistleblowing may even prove cost effective to the institution (Near and Miceli, 1985; Miceli et al., 2008). Despite this, and though improvements have been made (Freestone et al., 2006), whistleblowing is still not ubiquitous (DesRoches et al., 2010). Unfortunately, this hesitation to report poor practice is often warranted, as whistleblowers face reprisal including bullying, negative emotional/physiological effects, and even job loss (McDonald and Ahern, 2000; Jackson et al., 2014; Baucus and Dworkin, 1994; Matthiesen et al., 2011; Peters et al., 2011). Health employees appear cognizant of these dire consequences and often refuse to whistleblow for fear of reprisal and an apprehension in the efficacy of reporting (Attree, 2007; Kingston et al., 4; Black, 2011). As a consequence, significant research and legislation has sought to curb such reprisal, protecting the whistleblower (Ramirez, 2007; Eaton and Akers, 2007; Callahan and Dworkin, 1992; Bolsin et al., 2005; Faunce, 2004; Watson and OConnor, 2015). The implication being that if whistleblowing increases, bad practice would be excised and patient care would benefit.

In this work, we call into question the *a priori* assumption that internal whistleblowing (for example, to workplace colleagues, line managers or administrators) always augments patient care. We hypothesize that, given resource constraints and process inefficiencies, patient care can actually be improved if employees are somewhat hesitant to whistleblow. To test our hypothesis, we develop an agent-based model which considers the effects of whistleblowing on patient care. The employment of mathematical and computational modelling to investigate policy improvements has long been employed in many sectors, but, to date, has been underutilized in the healthcare arena (Pitt et al., 2015). Applying such tools permits us to validate hypotheses without significant cost, expediting policy conversation.

Recently, research has begun to focus on how the complex nature of healthcare institutions affects whistleblowing rates (Mannion and Davies, 2015; Jones and Kelly, 2013). The dynamics of institutions have been shown to significantly affect employee behaviour. Whistleblowing rates differ by nation states (Miceli and Near, 2013; Tavakoli et al., 2003; Trongmateerut and Sweeney, 2012), workplace culture (Mannion et al., 2005), and organization topography (King, 1999; Barnett, 1992). Given all the moving parts of an institution, it has been argued that culpability for the lack of whistleblowing cannot be placed solely on the healthcare worker (Dekker and Hugh, 2014). However, while research and governmental inquiries have provided insights into the complex nature of healthcare institutions and the cultural factors which inhibit whistleblowing, surprisingly little empirical work has considered how internal institutional processes could actually inhibit the benefits of whistleblowing. For example, staff surveys show that whistleblowing rates are correlated to the perceived efficacy of, and institutional response to reporting (Skivenes and Trygstad, 2015, 2010). Whilst nefarious causes for poor efficacy of internal reporting are a sad reality (e.g. the ignoring of complaints by upper management (Jones and Kelly, 2014a)), here we analyse a less malevolent, but more pervasive cause — organizational resource limitations and processing inefficiencies. We demonstrate that if

resources are limited or the processing of whistleblowing reports is inefficient, then patient care can actually be improved by infrequent whistleblowing and a lack of informational transparency.

Obviously acquiescing to a future of complicity and reduced transparency is objectionable. As such, we conclude the work by discussing mechanisms around this conundrum. If resource/efficiency constraints undermine the utility of employees' raising concerns, then we recommend a reduction of whistleblowing and the inclusion of softer mechanisms for internal adjustment of health care practices (Jones and Kelly, 2014b). For example, there is evidence to suggest that, whilst employees may not always formally blow the whistle, many do request changes to poor practice in less formal ways (Jones and Kelly, 2014a; Moore and McAuliffe, 2010), such as private peer to peer communication (Asghari et al., 2010). While softer mechanisms are inherently less robust than more formal internal investigations, we demonstrate that (in some contexts) a mixture of whistleblowing and softer assessments can provide better patient care than frequent formal inquiries.

8.3 Model Overview

Here we model the utility of informational transparency and whistleblowing in the healthcare sector. The healthcare institution is comprised of the following variables:

- Patients
- Healthcare workers
- The good and bad practices employed by the healthcare workers, which affect patient care.

The model is initialized with 10,000 workers, where 50% employ good practice (GP) and 50% employ bad practice (BP). The game is partitioned into 10,000 rounds. During a round of play the following events occur in order:

8.3.1 Round

1. With some likelihood i (informational transparency), a healthcare worker receives information about a random co-worker's practice.
2. If the worker believes the practice is bad, they whistleblow and demand an internal inquiry with probability w (propensity to whistleblow).
3. When a whistleblowing report is filed, then the inquiry is added to the investigation list.
4. Whistleblowing inquiries are resolved (see Section 8.3.2)
5. Finally, patients receive care. The patients receive care equal to the fraction of good practice in the institution.

8.3.2 Filing and Resolving Whistleblowing Inquiries

When the whistle is blown on a practice, a report is filed and added to the list of current investigations within the institution. Each investigation takes X = 100 rounds to finish. We presume the inquiry is 100% accurate, and if the inquiry is the result of bad practice, then the practice is immediately removed from the institution and replaced with good practice.

The investigation list contains all the practices currently under investigation. Each inquiry has a counter representing the number of rounds left until the investigation is complete. For example, if three practices are currently under investigation, then the list might appear as:

- Practice id #145 — 45 rounds left
- Practice id #2395 — 3 rounds left
- Practice id #89 — 100 rounds left (just added to the list)

After each round of play, a round is subtracted from each practice on the list. When the counter reaches zero for a practice, it is resolved, and the practice is set to good, improving patient care.

8.3.3 Analysing the Results

After 10,000 rounds of the game, the proportion of good practices in the institution is calculated over all rounds. This represents the average patient care at the institution. The game can be run for different parameter values and the best model for patient care is analysed.

8.4 Model 1: The (Utopic) Baseline

Here we consider how patient care is affected by both informational transparency within the institution ($i = [0, 1]$) and the propensity to whistleblow ($w = [0, 1]$). If $i < 1$, then there is a lack of informational transparency in the institution — workers do not always know the practices of their colleagues. If $w < 1$, then workers are somewhat complicit — when they witness a bad practice, they do not always blow the whistle. We consider patient care over all possible values of i and w, analysing how different values of transparency mix with varying rates of whistleblowing to affect patient care.

8.4.1 Results

Figure 8-1 represents patient care over a parameter sweep of i and w. The colour shade depicts patient care. Clearly, patients receive the best care when information is transparent ($i = 1$) and when employees are not complicit ($w = 1$).

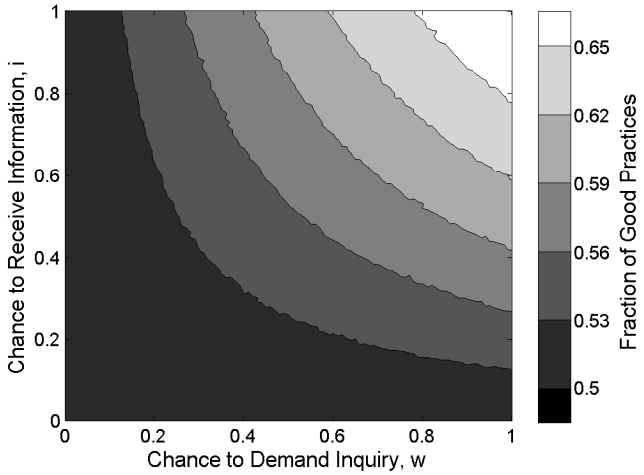


Figure 8-1: Patient care over a parameter sweep of informational transparency (i) and whistleblowing rate (w).

8.4.2 Discussion

This result offers defence to the presumption that, in an ideal world, without resource limitations, informational transparency and whistleblowing are both public goods. Patient care is best where there is informational transparency and a lack of complicity.

8.5 Model 2: When Patient Care is Improved By Complicity and Obfuscation of Information

The previous model presumes that the time it takes to review an inquiry is independent of the number of reviews currently being processed. This is not a practical assumption since resources are finite. If resources are committed to looking into practice A, it may delay the inquiry into practice B.

Here, we consider the repercussions of finite resources on patient care. Whenever someone whistleblows and adds a new practice to the inquiry list, a 3% increase in time is added to each inquiry. Since it takes 100 rounds to resolve an inquiry, a 3% increase means that 3 rounds are added to each inquiry in the list.

For example, presume the current inquiry list is:

- Practice id #145 — 45 rounds left
- Practice id #2395 — 3 rounds left
- Practice id #89 — 100 rounds left

If a healthcare worker blows the whistle on another practice, then the resulting list would appear as:

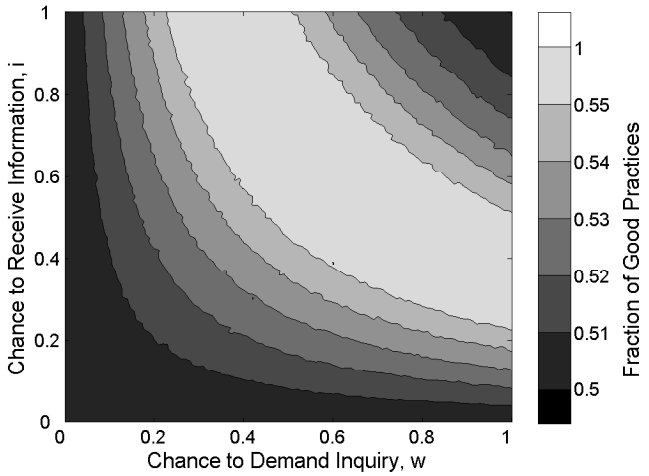


Figure 8-2: Average patient care after 10,000 rounds when inquiry time is increased by 3% per additional review.

- Practice id #145 — 48 rounds left
- Practice id #2395 — 6 rounds left
- Practice id #89 — 103 rounds left
- Practice id #9845 — 109 rounds left

Three rounds are added to each existing inquiry. Furthermore, the new inquiry (id #9845) starts with 100 + 9 rounds, because 3 rounds are added for each pre-existing inquiry already in the list.

8.5.1 Results

As a result of adding a 3% delay, Figure 8-2 illustrates that informational transparency and frequent whistle-blowing no longer lead to optimal patient care. If information is transparent ($i = 1$), then patient care is best if workers only whistleblow at a rate of approximately 35-45%. Further, if policy dictates that every worker must whistleblow whenever they witness a bad policy ($w = 1$), then it is best for patient care if information is partially obfuscated with workers only receiving information with a frequency of 35-45%.

8.5.2 Discussion

Model 2 shows that whilst whistleblowing leads to the eventual improvement of a practice, where there are finite resources it also delays investigation into other practices. This prolongs the amount of time these potentially damaging practices are in circulation. As a result, occasionally not learning about a bad practice

(i.e. lack of transparency), or not whistleblowing (i.e. complicity) may lead to improved patient care through more timely resolution of existing investigations.

It is worth noting that delaying the resolution of an investigation by 3% is quite small. If we presume that the institution's budget is static, then the percentage would be 100%. Further, we are presuming that the resources required to maintain an impressive efficiency of 3% does not negatively impact the resources employed for direct patient care. Both assumptions are fairly generous, but even at 3% it is still not efficient enough to overcome the utility of complicity or informational obfuscation. Appendix B.1 demonstrates that the utility of complicity and informational ignorance significantly increases if 5% or 10% are added to review times (though this is still quite generous).

In the health care sector, significant research, regulation and policy imperatives have sought to increase whistleblowing rates and informational transparency. Here we demonstrate that increasing whistleblowing rates may not always prove beneficial as this can delay organisational learning and correction of existing bad practice. Even in a world where whistleblowers do not fear sanctions, it may not be in the institution's or patients' best interest to always advocate whistleblowing. If resources are even slightly constrained, some amount of complicity may improve patient care.

8.6 Model 3: More Efficient Reviews

In this model we demonstrate how increased efficiency in processing inquiries empowers whistleblowing and informational transparency. Previously, each inquiry took 100 rounds to process. Now, we reduce the number of rounds per inquiry to 50. An additional 3% is still added to each inquiry when a whistle is blown, but since the initial inquiry is 50, only 1.5 rounds are added.

8.6.1 Results

Figure 8-3 illustrates that it is now best for patient care if information is transparent and workers always whistleblow. By reducing processing time, the benefits of complicity dissolve — after 10,000 rounds it is best if workers whistleblow when they witness bad practice.

8.6.2 Discussion

In the experimental literature, efficacy of whistleblowing systems is correlated to the propensity to whistleblow (Skivenes and Trygstad, 2015, 2010). Practitioners in institutions with highly effective formal review processing are more likely to whistleblow compared to those in less effective institutions. Models 2 and 3 demonstrate that workers in both types of institutions could be operating in such a manner as to increase the quality of patient care. In Model 2, we showed that increased processing time reduces the benefits of a whistleblowing system to patient care. However, if processing time is halved, suddenly frequent whistleblowing and informational transparency performs best.

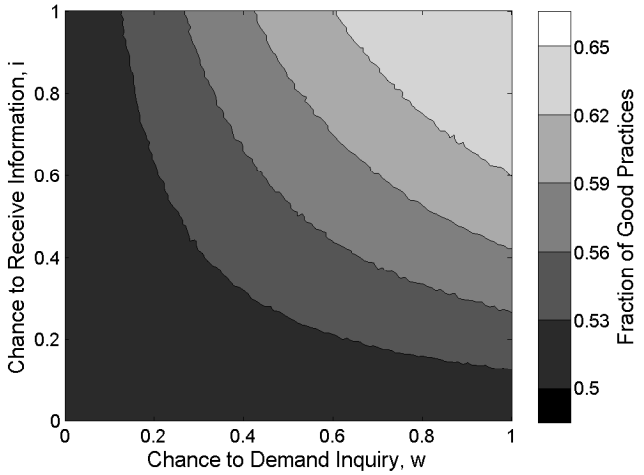


Figure 8-3: Average patient care after 10,000 rounds when the initial review time is 50 rounds.

8.7 Model 4: The Consequences of Time

Why does halving the processing time of an inquiry alter the optimal rate of whistleblowing? To investigate this question, we consider how patient care changes over time. Thus far, we have considered patient care over 10,000 rounds. Now, we extend Model 2 and Model 3 to 50,000 rounds and analyse how different whistleblowing strategies affect patient care over time.

8.7.1 Results

Running Model 2 for 50,000 rounds, rather than just 10,000 yields a very different result. Suddenly, whistleblowing and transparency provide the best patient care. After 50,000 rounds, the graph of patient care looks quite similar to Figure 8-3 (see Appendix B.2).

Figure 8-4 illustrates why. The solid lines depict the proportion of good practices in the institution over time. The chance of attaining information is held static at 100% ($i = 1$), and patient care is measured against different whistleblowing strategies. The square line represents an institution where no one whistleblows ($w = 0$); consequently the number of good practices never increase. The other lines represent different whistleblowing strategies, ranging from 30% ($w = 0.3$), to always whistleblowing ($w = 1$).

For the first 10,000 rounds patient care is hindered by constant whistleblowing. Some complicity ($w < 1$) outperforms unconditional whistleblowing ($w = 1$). Since high whistleblowing rates result in frequent inquiries, the resolutions of existing inquiries are delayed and patient care suffers. However, although frequent whistleblowing delays inquiries, once the inquiries begin to resolve, a higher number of practices are altered. At some point, frequent whistleblowing begins to catch up and eventually overtakes less frequent

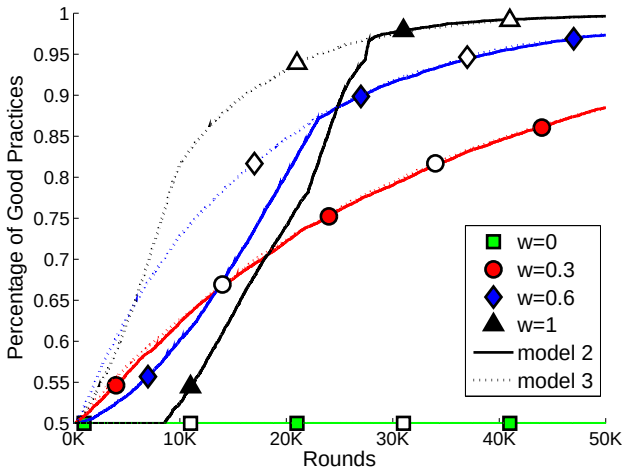


Figure 8-4: Patient Care over 50,000 rounds for different whistleblowing and efficiency rates. The chance to receive information is held static at 100% ($i = 100$). **Solid lines:** Initially, reviews take 100 rounds (Model 2). **Dotted Lines:** Reviews take 50 rounds (Model 3). **Squares:** No one ever blows the whistle, $w = 0$. **Circles, Diamonds, and Triangles:** w is 30%, 60%, and 100% respectively.

whistleblowing. In Figure 8-4, for the first 10,000 rounds, 30% whistleblowing (circle line) outperforms 60% (diamond) and 100% (triangle) whistleblowing. However, after round 13,438, $w = 60\%$ overtakes $w = 30\%$ — it is better for patient care if whistleblowing is more frequent. At round 18,199, a refusal to be complicit ($w = 100\%$) passes $w = 30\%$, and at round 24,800, $w = 100\%$ passes $w = 60\%$.

For Model 3, where inquiry processing only requires 50 rounds, we showed that after 10,000 rounds whistleblowing outperforms complicity. Importantly, the mechanism behind the behaviour is the same as when inquiries require 100 rounds. Consequently, for Model 3 we would expect there exists a time before which complicity is advantageous.

The dotted lines in Figure 8-4 depict Model 3 over 50,000 rounds. Clearly, unconditional whistleblowing ($w = 1$) outperforms all other strategies well before 10,000 rounds. Though it is difficult to see in the graph, by round 890, $w = 100\%$ surpasses $w = 30\%$. This is in contrast to 13,438 rounds when inquiries take 100 rounds. Further, $w = 100\%$ outperforms $w = 60\%$ at round 5669. Again, this is in contrast to 24,800 rounds when inquiries take 100 rounds.

8.7.2 Discussion

Since, given enough time, it is always eventually beneficial to whistleblow without complicity ($w = 1$), is it important to consider the early stages of institutional improvement, where complicity or lack of transparency generates better patient care? Why not accept that change comes at a cost, and while the institution's performance may be hindered for a short period, in the end transparency and whistleblowing lead to the best

patient care?

Evidence suggests that time may be a luxury that cannot be afforded. It is known that workers who experience delays in the process of handling and investigating concerns are often prone to absence from work due to physical and mental ill health (Francis, 2015). This all costs additional money which creates a feedback loop, negatively impacting patient care and staff morale (Francis, 2015). Additionally, as previously mentioned, a 3% increase in review time is quite generous. If the percentage is increased to 5% or 10%, then it requires even longer before unconditional whistleblowing, $w = 1$, outperform a fair amount of complicity, $w = 0.3$ (see Appendix B.3).

However, in certain circumstances it could very well be worth the slow start in order to increase good practice at a later date. By decreasing the length of inquiries by half (100 to 50 rounds), it takes much less than half the time for whistleblowing to outperform complicity. There is, at least in this simple model, an exponential benefit for decreasing the amount of time it takes to process inquiries.

In this model, we explained why institutional differences in the efficacy of formal reviews can lead to divergent strategies for increasing patient care. In each institution, there is always a period of time where some amount of complicity outperforms certain whistleblowing; however, linear cuts in processing time lead to exponential cuts in the time complicity outperforms whistleblowing. As a consequence, whilst it might be feasible to accept the slow start to whistleblowing in an efficient institution, it might prove less feasible if inquiries take longer to process. This provides further evidence that the complex nature of institutions should be studied prior to recommending or implementing a strategy of carte blanche whistleblowing.

Lacking Veridical Knowledge

Finally, in this section we analysed the effects of whistleblowing rates whilst presuming that information was freely transparent ($i = 1$). However, identical results hold if we presume individuals always whistleblow ($w = 1$), and test informational obfuscation (i.e. $i = 0; 0.3; 0.6$) against transparency ($i = 1$). In efficient institutions, informational transparency quickly usurps institutions with hidden information. However, institutional inefficiency dramatically lengthens the period of time where opaque institutions outperform transparent institutions.

8.8 Model 5: Soft, Unmonitored Advice Versus Whistleblowing

Resigning ourselves to complicity and reduced transparency seems an unacceptably risky strategy. It is in precisely these types of environments where malevolent practices or simply unimpeded incompetence can lead to catastrophe. In this section, we discuss how to circumvent the problems introduced by inefficiencies. We do this by analysing a mixed strategy of whistleblowing and less monitored, but less resource intensive solutions for altering practice.

Up until now, when a worker was presented with bad practice, they were only able to decide whether or

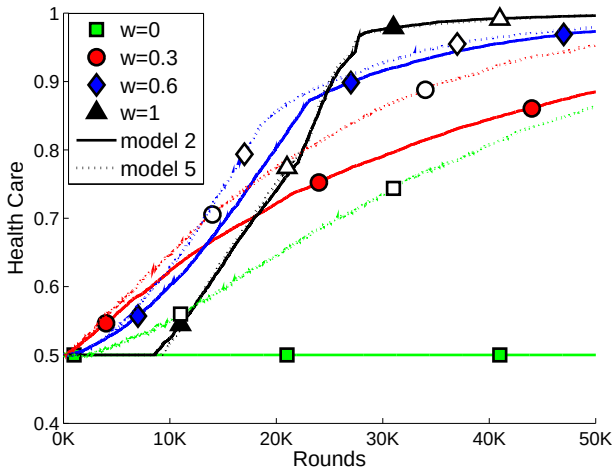


Figure 8-5: Comparison of patient care with and without soft advice. The chance to receive information is held static at 100% ($i = 1$). Inquiries take 100 rounds with a 3% dependency. **Solid lines:** No soft advice; $s = 0$ (Model 2). **Dotted Lines:** Soft advice, $s = 1$ (Model 5).

not to whistleblow. Here we introduce a softer mechanism for altering practice. Rather than whistleblowing, the employee may advise a workplace colleague in an attempt to alter their behaviour without formal internal investigation.

To model this we add the propensity of an individual to employ soft advice, s. As before, when an employee witnesses bad practice, they whistleblow with probability w. However, if the employee does not whistleblow, then they offer advice with probability s. If the agent offers advice, then their colleague may update their behaviour based on the informal advice.

One of the concerns for permitting softer behavioural changes is their lack of transparency. While internal investigations may take time, they are more transparent than permitting workers to share information without moderation. What happens if the advice given to the bad practising agent is not beneficial, or even worse, if it further impairs the practice?

To analyse the inherent risk of soft advice, we presume that when someone informs another of bad practice, their advice can either 1.) improve, 2.) worsen, or 3.) not affect the practice. Thus far, each good practice has increased patient care by 1, and each bad practice was represented as a zero. Now we add the ability to negatively impact patient care with a -1. If a practice is improved, then, as in the formal review, the bad practice (0 or -1) is switched to a good practice $= 1$. If a bad practice is worsened, then the practice is switched from a 0 to a -1 (or if it is already a -1, it remains a -1). Consequently, the practice actually hinders patient care. Finally, soft advice can be ignored, and the practice retains its current value. We presume the chance of each output is equally probable ($1/3$ to improve, worsen, and not affect). This is in contrast to whistleblowing which, in our model, will always result in a positive shift.

8.8.1 Results

Figure 8-5 compares Model 2, where soft advice is not permitted (solid lines), with an institution where workers always softly advise, $s = 1$ (dotted lines). The chance to receive information is held static at $i = 1$ (i.e. full transparency). Patient care is measured for different whistleblowing strategies. The y-axis is slightly altered compared to previous graphs. Since, in prior models, good practice = 1 and bad practice = 0, the fraction of good practices was the measure of patient care. Now, practices are in the set $\in [-1, 0, 1]$. Consequently, health care is measured by the average value of all practices.

Figure 8-5 shows that it is beneficial to promote an environment where staff offers unmonitored advice, even if the advice could worsen the practice. If practitioners whistleblow with a 30% or 60% frequency, then soft advice always outperforms practitioners who do not softly advise. This is demonstrated by the fact that the dotted line is always higher than the solid line for both $w = 0.3$ (circles) and $w = 0.6$ (diamonds).

If workers always whistleblow ($w = 1$), then soft advice offers no benefit. This is as expected, since an employee only softly advises if they do not whistleblow, but since they always blow the whistle, the potential advantages of soft advice are never witnessed. Finally, the dotted square lines illustrates that if soft advice is utilized without any whistleblowing ($w = 0$), then practices improve, though not as much as a mixture of whistleblowing and soft advice.

8.8.2 Discussion

Despite the trepidation of unmonitored advice, in this simple model, it is almost always beneficial to permit soft advice. However, it is important to note that in Figure 8-5 there is an equal chance that advice will help, harm, and not affect practice. If the likelihood of harmful advice is increased, then the benefits of soft advice may diminish or be eliminated (see Appendix B.4). This further establishes the need to consider the dynamics of the institution prior to instantiating whistleblowing policy. While it seems unlikely that professional advice would prove more likely to harm than help, such mechanisms play a significant force in selecting the best policy.

Importantly, in Figure 8-5, it is the mixture of whistleblowing and soft advice which performs best. Although unconditional whistleblowing eventually surpasses strategies with soft advice, it requires more time. When the whistleblowing rate is 60% (dotted, diamond line), unconditional whistleblowing ($w = 1$) only surpasses it when almost all practices are good. Thus, waiting for the benefits of unconditional whistleblowing may not be worth delaying the immediate improvement to health care given a mixture of soft advice and whistleblowing. If inefficiencies in processing cannot be overcome, soft advice may augment whistleblowing generating improved patient care.

8.9 General Discussion

In this work, we have argued that subtle and small perturbations in a healthcare system may lead to significant alterations in the best policy for maximizing patient care. By adding 3% interdependence between review processing time, we found that complicity and a lack of transparency can, paradoxically, benefit patient care (see Model 2). Whilst the benefits of complicity do not hold in the long term (see Model 4), for a given institution it may be long enough to prove salient, altering the best strategy for patient care. However, if inquiry time is reduced (see Model 3 and 4), then this effect may prove small enough that the initial penalties for whistleblowing may be weathered and the future benefits of a transparent institution enjoyed. This may help explain why whistleblowing rates are correlated to the perceived efficacy of, and institutional response to reporting (Skivenes and Trygstad, 2015, 2010).

We further analysed whether softer mechanisms for updating practices are beneficial to the health care system. Whilst unmonitored advice seems risky, we showed that a mixture of whistleblowing and soft advice may optimize patient care. While the fidelity of internal institutional whistleblowing inquiries is valuable, it is cost prohibitive. Cheap but less accurate measures for updating practices mixed with highly accurate, but costly internal inquiries leads to excellent patient care.

Clearly, this is a simple model; however, this work is not trying to emulate a working health care system. Instead, we wish to demonstrate how small perturbations in the system can lead to significant equilibrium shifts. We only added 3% delays to whistleblowing inquiries, and as a consequence complicity and obfuscated information were actually beneficial to patient care. We could have included several other roadblocks to resolving inquiries. For instance, given a static budget, whistleblowing reviews might actually hinder the care patients receive. If the institution is spending money on reviewing, it cannot spend money on additional patient needs. Further, if a policy or guideline is created because of care failings in one institution, should other well-operating institutions be forced to include the new policy? The dexterity required to implement new policies may be non-trivial, potentially creating more problems than they solve. There are a number of ways in which the costs of whistleblowing may prove inhibitive, we explained one of these.

Significantly more research and theory is needed to diagnose a complete understanding of how the utility of whistleblowing and information transparency relates to the complexities of health care institutions. The present work calls into question the *a priori* belief that whistleblowing and transparency improve patient care, and suggests that given resource constraints, a mixture of soft and formal reporting may provide the best patient care. We hope continued research will aid in explaining how organizational complexities affect the utility of whistleblowing, leading to improved patient care. Further, we join the recent call (see Pitt et al., 2015) for further utilization of the relatively untapped resource of modelling techniques into expediting and validating healthcare policy conversations.

9

Conclusions

> *Nothing is so difficult as not deceiving oneself.*
>
> Ludwig Wittgenstein, *Culture and Value*

> *I was never entirely certain why he found my mental absences so irritating — whether he thought I was being wilfully obtuse to annoy him or whether he felt I was unreasonably cheating hardship by failing to notice it — but I made a private pledge to remain alert and fully conscious for a while, so not to exasperate him.*
>
> Bill Bryson, *A Walk in the Woods*

9.1 Thesis

A veridical understanding of the world is typically invaluable in navigating complex environments. However, in a variety of contexts, individuals do not employ veridical knowledge during action selection. This is true even when such information is freely available. In this work, I have offered an explanation for this phenomenon. My thesis is that there are a surprising variety of contexts where (at least partially) biased knowledge is adaptive. Throughout this dissertation I offered explanations for a variety of open scientific questions by demonstrating the benefits of possessing incorrect information. I showed that lacking veridical knowledge can be adaptive in maintaining cooperation, selecting actions in unpredictable environments,

investing in markets, and operating within an inefficient institution. In Chapter 2, I argued that adaptive ignorance is, at minimum, a necessary condition for the evolution of self-deception. If, as many evolutionary theorists seem to believe, adaptive ignorance is a sufficient condition for self-deception, then this dissertation argues at the breadth of environments where self-deception is advantageous. If adaptive ignorance is only a necessary requirement for self-deception, then this dissertation demonstrates that one of the environmental building blocks for self-deception is (at least theoretically) pervasive. In this final chapter, I summarize my findings, discuss limitations and potential criticisms, and propose future work.

9.2 Chapter Summaries

This dissertation argues that there are many contexts where it is advantageous to lack veridical knowledge. Chapter 3, argues that the value of ingroup cooperation may outweigh the costs of acting honestly. Even when information is free, cooperation is better maintained by siding with false, social beliefs. This may explain why individuals use error-prone social information in lieu of free, veridical information, when deciding whether to act prosocially (Sommerfeld et al., 2008).

In Chapter 4, I analysed a potential advantage of the impact bias. The impact bias refers to the human propensity to exaggerate affective forecasts (Wilson and Gilbert, 2013; Wilson et al., 2000; Schkade and Kahneman, 1998; Wilson et al., 2003a; Wilson and Gilbert, 2003; Gilbert et al., 1998). Whilst the impact bias can be harmful (Janz et al., 2007; Ruby et al., 2011; Dillard et al., 2010), recently there has been a call to discover whether the bias can provide some benefit (Miloyan and Suddendorf, 2015; Marroquín et al., 2013). Extending the work on Error Management Theory (see Johnson et al., 2013; Haselton and Nettle, 2006; McKay and Efferson, 2010), I demonstrate that mispredicting future affective experience can be beneficial by compensating for noisy information.

Chapter 5 shows that blind trust in others and the rejection of unfair offers (i.e. costly punishment) is revenue maximizing given partial information and partner choice. When individuals do not know whether a potential partner is trustworthy, they tend to blindly trust the other despite the risk of deleterious deals (Berg et al., 1995). Interestingly, when they do know that a partner will offer profitable deals, they often reject the offer unless it is fair (Henrich et al., 2001; Marlowe et al., 2010). I argue that such behaviour is adaptive in a market where information about partners is incomplete. Over the last couple of decades, evolutionary theorists have analysed the contexts where costly punishment is adaptive (Debove et al., 2015; Barclay and Stoller, 2014; Huck and Oechssler, 1999; Rand et al., 2013). This work represents a parsimonious explanation for why individuals costly punish. Whilst it does not explicitly argue that lacking veridical knowledge is beneficial, it represents a prerequisite for Chapter 7, costly punishment can be adaptive.

Whilst significant research has focused on explaining the adaptive nature of costly punishment, few have explained the heterogeneity of individual strategies (Fischbacher et al., 2013; Manapat et al., 2012). Chapter 6 extends the model in Chapter 5, arguing that mixed strategies of costly punishment can be adaptive. If knowledge of the optimal strategy is too costly to acquire, or is impeded by some other force (e.g. the

second-order free rider problem, see Sasaki et al., 2015; Boyd and Richerson, 1992), then it can be adaptive for an ignorant population to attempt a variety of strategies.

Chapter 7 demonstrates that a variety of strategies can neutrally drift into a population playing the Ultimatum Game. Previous attempts to explain the heterogeneity of human strategies have focused on uncovering some variable which explains the exact frequency distribution of strategies in a population (Manapat et al., 2012; McNamara et al., 2009b). Chapter 7 offers a more parsimonious explanation. If rejecting unfair offers is advantageous (see Chapter 5), then other strategies are free to drift into the population. As long as a sufficient subset of the population possess the optimal strategy of costly punishment, then suboptimal strategies go unpunished. There is no need to define a function for the frequency distribution of a population when neutral drift is permitted.

Finally, Chapter 8 demonstrates that small inefficiencies in a healthcare organization engender the partial obfuscation of information. In a perfect world, transparency of information is always better for patient care. This has led to a call for policy advocating pervasive internal whistleblowing (Ramirez, 2007; Eaton and Akers, 2007; Callahan and Dworkin, 1992; Bolsin et al., 2005; Faunce, 2004; Watson and OConnor, 2015). However, research is only beginning to consider how the complexity of healthcare institutions affects the institution's ability to abide policy (Mannion and Davies, 2015; Jones and Kelly, 2013). To date, theoretical simulations of the consequences of policy are all but non-existent in the healthcare arena (Pitt et al., 2015). We demonstrate that obfuscation of information can lead to improved patient care given small inefficiencies in the institution. Awareness of this problem will hopefully lead to the integration of practicalities with policy, improving patient care.

9.3 Limitations and Future Work

Existence Proofs do not Prove Existence: Empirical Testing Required

The work presented here has been theoretical in nature. It stands as an existence proof for the idea that lacking veridical knowledge can be advantageous in a variety of contexts. Further, it demonstrates that several open questions regarding human behavioural biases can be answered by seriously considering the notion that veridical information can be detrimental to action selection. Of course, this work does not prove these answers to be the cause of human behaviour. This work represents theory building and is preliminary in understanding the evolutionary and cultural constraints on human cognition. Further empirical testing is required to help uncover the underlying causes of human behaviour. For example, as shown in Chapter 4, the impact bias is advantageous in noisy environments due to Error Management effects. Importantly, this does not prove that there was selective pressure for the impact bias. However, given the current state of empirical evidence, it is a plausible explanation.

In the future, empirical testing should be conducted in order to compare existing explanations for the impact bias. Currently, there are two theories offering explanations for the bias. My own theory is similar

to the recent hypothesis that the impact bias is advantageous for motivating decision-making (Morewedge and Buechel, 2013; Greitemeyer, 2009). Another theory hypothesizes that the measured effect is mostly a regression to the mean, and that individuals are poor at predicting that, over time, all affective experiences shift to innocuity (Levine et al., 2012). To my knowledge, no one has explicitly, empirically tested these two theories against each other. Such an empirical test would advance theory.

Is There Plasticity Within an Individual Lifetime? Empirical Testing Required

Most of this work has been argued over an evolutionary time-frame. As previously mentioned, this is not because my arguments are inherently evolutionary. Evolutionary algorithms are particularly useful in discovering advantageous strategies in frequency dependent environments (Alexander, 2009). Simulated evolution is simply a learning mechanism and can be applied across lifetimes or within a lifetime.

In the future, empirical research should diagnose whether individuals are sensitive to the dynamics presented in this work over the course of one life-time. For example, in Chapter 4 I demonstrate that the impact bias is advantageous in unpredictable environments. This bias could be learned over an evolutionary time-frame, or within one's lifetime. A future, empirical study could consider whether individuals adjust affective forecasts based on the predictability of the environment.

How Does Self-Deception Relate to Lacking Veridical Knowledge?

Another open question remains, how does self-deception relate to contexts where it is adaptive to lack veridical knowledge? In Chapter 2, I argue that adaptive ignorance is, at minimum, a necessary condition for the evolution of self-deception. Further, for some, lacking veridical knowledge and self-deception represent one and the same phenomenon.

If you believe the two ideas are identical, then this question is moot and the present work demonstrates the ubiquitous evolutionary feasibility of self-deception. However, if lacking veridical information is a necessary requirement of self-deception, what other variables mediate selective pressure for self-deception? Again, in the future, theory and empirical data need to better align. At the very least, the present work demonstrates that one of the necessary conditions for self-deception is (at least theoretically) pervasive.

9.4 Common Environmental Features for Adaptive Ignorance

When I first began researching for this work, I intended to focus on uncovering the common environmental features which lead to an advantage for self-deception and lacking veridical knowledge. However, in undertaking the task, I realized the literature was too nascent for validating a general theory for adaptive ignorance. As a consequence, the present work does not attempt to categorize the types of environments which generate pressure for lacking veridical knowledge. Rather, this work represents the data gathering stage. By discovering new contexts where it is beneficial to lack veridical knowledge, more data points are

available when attempting to form a generalized framework explaining the common environmental features required for adaptive ignorance.

Having said that, I think it important to include a brief discussion on the current state of literature in attempting to categorize the environmental features necessary for the evolution of self-deception, misbelief, and cognitive biases. As mentioned in Chapter 2, several have attempted to categorize the necessary environmental features. Haselton et al. (2015) suggest that cognitive biases are caused by either heuristics, error management effects, or experimental artefacts. McKay et al. (2009) argue that the only advantageous phenomena which require a misbelief are the health benefits incurred from optimism (e.g. Taylor and Brown, 1988; Creswell et al., 2007; Sharot, 2011). Chance and Norton (2015) summarize the adaptive nature of self-deception, noting that self-deception may be beneficial in order to deceive others, garner social benefits, or to increase psychological well-being. In contrast, von Hippel and Trivers (2011a) argue that self-deception is only beneficial for deceiving others (though garnering social benefits falls under this banner, see von Hippel and Trivers, 2011a).

Adaptive Ignorance Requires Faulty Decision-Rules

I believe Steven Pinker helps clarify the confusion. Pinker (2011) argues that, when possible, veridical knowledge should always be employed when making a decision. To make his point, Pinker dissects the decision-making process into two parts. First, there is the information used in making the decision. This information may or may not be accurate. Second, there is the decision-rule which uses the information, resulting in an action. If the amalgam of one's knowledge and decision-rule result in poor actions, then one may improve action selection by altering either the information or the decision-rule. Pinker argues that, if given the choice, it is always better to alter the decision, retaining veridical knowledge.

Pinker's delineation between decision-rules and information offers an explanation for many of the instances where it is adaptive to ignore veridical information. In situations where the decision-rule cannot or will not be altered, biasing information may prove advantageous. For instance, the less-is-more effect occurs when more information generates worse decisions (Czerlinski et al., 1999; Todd and Gigerenzer, 2000). An example of this is the recognition heuristic (Goldstein and Gigerenzer, 1999). When American and German students are asked whether San Antonio or San Diego has a larger population, Germans perform better than Americans (Goldstein and Gigerenzer, 2002). This is because the students use the recognition heuristic as their decision rule — they select the city that is most recognizable. Interestingly, the rule performs worse with more information, since more cities are recognized (Gigerenzer and Goldstein, 1996). Since the German students possess less information, and the option they recognize is highly correlated to the largest population, they outperform American students (Gigerenzer and Brighton, 2009, for a nice review of Gigerenzer's work see Boudry et al., 2015).

If an American student is attempting to improve their score, they could either 1.) forget information, and employ the recognition decision-rule, or 2.) change their decision-rule. Clearly, Pinker (and most others) would claim option 2 the better. Interestingly, the idea that fallible decision-rules create scope for adaptive

ignorance may explain much of the work presented here.

Categorizing the Present Work

Consider the findings in Chapter 3. In environments with high value-homophily, it is adaptive not to use veridical information. However, it is not harmful to possess veridical knowledge, it is only harmful if one employs it when deciding whether to cooperate with another. Since it is suboptimal to both possess and employ veridical information, the agent can improve their action in one of two ways. They could either: 1.) always use the information they possess (decision-rule), but only seek out social information (bias information), or 2.) always seek out veridical information (do not bias information), but know when to use it (alter decision-rule).

In Chapter 8, under certain contexts a mixture of informational obfuscation and hesitation in whistle-blowing can improve patient care. If the decision-rule is to always blow the whistle, then it is advantageous to reduce information transparency (bias information). However, if the decision-rule is altered to consider instances where the whistle should not be blown, then informational transparency is again optimal.

In Chapter 4, I presume that individuals define the value of an action based on previous experience — this is their decision rule. Since the world is unpredictable, it is advantageous to bias the values of their previous experiences. Again, this can be done in one of two ways. An individual could retain the decision-rule but alter their memory of past experience — as seems to be the case with human subjects (see Morewedge et al., 2005; Fredrickson et al., 1993). Alternatively, an individual could realize that, in noisy environments, it is important to bias their affective experience. In this case, the individual could accurately recall their affective experience, but subsequently bias it prior to action selection.

Finally, in Chapter 6, I show how altering a decision-rule changes the type of information that is required. If the decision-rule is to always costly punish offers below a minimal acceptable offer (MAO), then simulated evolution finds an advantageous MAO (see Chapter 5). However, this decision-rule may be limited due to other factors, such as the second-order free rider problem (Sasaki et al., 2015; Boyd and Richerson, 1992). In such situations, Chapter 6 demonstrates why mixed strategies might occur (Krasnow et al., 2015; Chen et al., 2014). A mixed strategy, even one where the individual is willing to accept detrimental offers in 50% of the games, can prove adaptive. If an agent cannot find the optimal MAO to use the best decision rule, an individual may play a mixed strategy since they do not need as much information to gain an advantage.

Categorizing Previous Work

Some previous attempts to categorize the advantages of lacking veridical knowledge can be simplified by the idea that adaptive ignorance requires faulty decision-rules. Haselton et al. (2015) argue that cognitive biases are generated by heuristics, error management effects, or experimental artefacts. These categories can be coalesced under the banner of faulty decision-rules. First, heuristics are *defined* as fallible decision-rules. Second, the cognitive biases caused by error management effects (such as those witnessed in Chapter 4) can

be removed by altering one's decision-rule (McKay and Efferson, 2010). If one is aware of error management effects, then (as explained above) the decision-rule can be altered, rather than the information. Finally, Haselton et al. (2015) suggest that some experiments lead to cognitive biases because they are far removed from natural environments. However, this is a similar argument to heuristics, and can be explained by altering decision-rules. Because individuals have never witnessed a particular experimental design, they use suboptimal decision-rules for action selection. If they knew the best decision-rule for acting in a particular experiment, they could perform optimally.

Importantly, there are exceptions where faulty decision-rules do not appear to cause adaptive ignorance. von Hippel and Trivers (2011a) argue that self-deception augments interpersonal deception. Individuals can exploit the limitations of deception detection by deceiving themselves. In another example, individuals recover from maladies at a higher rate, if they are overly optimistic (Taylor and Brown, 1988; Creswell et al., 2007; Sharot, 2011, also see McKay et al., 2009). The reason underlying this is still not understood. Finally, in Chapter 7, I demonstrated that ignorant strategies can drift into a population as long as the strategies of some subset of the population protect them from being punished.

Altering the Decision-Rule is Not Necessarily Feasible or Desirable

Pinker (2011) takes a hard stand in claiming that altering a decision-rule is always preferential to biasing information. Here, I disagree. Even if fallible decision-rules are culpable for many of the contexts generating adaptive ignorance, it may be neither feasible nor desirable to alter the decision process.

Just as temporal limitations constrain one's ability to procure information (Simon, 1972, 1991), deducing the best decision rule is likely non-trivial. It is well known that Bayesian inference in complex, natural environments is infeasible (Marshall et al., 2013a). This has led Fawcett et al. (2014) to categorize the types of environments which lead to simple, through occasionally vulnerable, decision rules. They argue that a variety of human biased heuristics are ecologically rational in environments of uncertainty, autocorrelation, heterogeneity, or state-dependence (Fawcett et al., 2014).

Even if Bayesian inference can be employed, in natural, complex environments, simple heuristics often outperform "rational" Bayesian updating (Lee et al., 2002) or multiple regression (Czerlinski et al., 1999). Whilst more complex learning mechanisms are terrific at mapping a function to data, they often overfit their predictions in a world that is always changing (Gigerenzer and Brighton, 2009). As a consequence simple heuristics can prove better at predicting the future compared to more statistical methods. So, while these heuristics may be vulnerable to permitting adaptive ignorance, in the long run, it might be better to retain the heuristic.

Finally, evolution is a blind watchmaker. If decisions can be altered either through the decision-rule, or via biasing perception, we cannot yet predict which avenue evolution will take. Further research into recognizing how the malleability of decision-rules affect adaptive ignorance may offer insight into the scaffolding of cognition.

9.5 Conclusion

When making decisions, humans bias and ignore veridical information in a variety of contexts. In this work, I have answered several open scientific questions by demonstrating that such biases can be adaptive. Lacking veridical knowledge can be beneficial in navigating cooperative societies, unpredictable environments, investment markets, and inefficient institutions. The value of these studies are multi-faceted. First and foremost, they offer ultimate explanations for previously confusing human behaviour. Second, they offer insight into a more general explanation of when we might expect individuals to ignore veridical information and self-deceive. Whilst significant work remains in deducing the forces underlying decision-making, this work demonstrates the ubiquity of environments where individuals may have stumbled upon the value of ignorance.

Appendix A

Appendix for Chapter 3

A.1 Social Norms

A.1.1 Model 2: Homophily

Thus far we have shown that the Judging norm (see Table 3.2) and homophily can increase the robustness of cooperation in the face of erroneous communication. Here we check whether this result holds for other social norms. We employ three norms, shown in Figure A-1.

With the Standing norm — Figure A-1(a) —, if a donor cooperates with the recipient, it always receives a good reputation. The only situation where an agent becomes ill-reputed is if it refuses to cooperate with a good recipient. When agents always share the same belief about an given agent, even if unjustified, the Standing norm has been shown to maintain cooperation (Ohtsuki and Iwasa, 2004). However, Takahashi and Mashima (2006) showed a Standing society is susceptible to invasion if error is disseminated like our model, where agents can have differing opinions regarding the same agent.

The Shunning norm (Figure A-1(b)) is the only norm other than the Judging norm which sustains cooperation when agents can hold different beliefs about the same agent (Takahashi and Mashima, 2006). With Shunning, an agent can only receive a good reputation if it cooperates with a good recipient. Finally, with Image Scoring — Figure A-1(c) — an agent receives a good reputation for cooperating, and a bad reputation for defecting. As a first order norm, it is well-known to fail to maintain cooperation (see Section 3.3.2).

Results

Figure A-2 depicts the fraction of cooperative actions performed utilizing three different norms for two values of homophily. In a society using the Standing norm without homophily — Figure A-2(a) — cooperation is vulnerable because ALLCs can perform better than DISCs and are subsequently vulnerable to invasion by

	Good	Bad
Coop	G	G
Defect	B	G

(a) Standing

	Good	Bad
Coop	G	B
Defect	B	B

(b) Shunning

	Good	Bad
Coop	G	G
Defect	B	B

(c) Image Scoring

Figure A-1: Definition of tested social norms. Each column represents whether the observer believes the recipient is good or bad. The first row is the reputation the observer will impart on the donor, if the donor cooperates. The second, is the reputational repercussions if the donor defects.

ALLDs (see Takahashi and Mashima (2006) for a discussion). Interestingly, in a society with homophilous interactions — Figure A-2(b) — cooperation is aided and more robust against invasion.

Figure A-2(c) illustrates the cooperative propensity of the Shunning norm without homophily. The Shunning norm has been shown to remain stable when $e = 0.025$ (Takahashi and Mashima, 2006). Here, we extend this research and show that, for certain levels of benefit (b), cooperation can remain stable in the face of erroneous communication. Furthermore, homophily extends this stability, stabilizing cooperation despite lower values for b, and higher error rates. Finally, we reproduce the well known result that Image Scoring cannot sustain cooperation — Figure A-2(e). Furthermore, homophily does not aid in maintaining cooperation — Figure A-2(f).

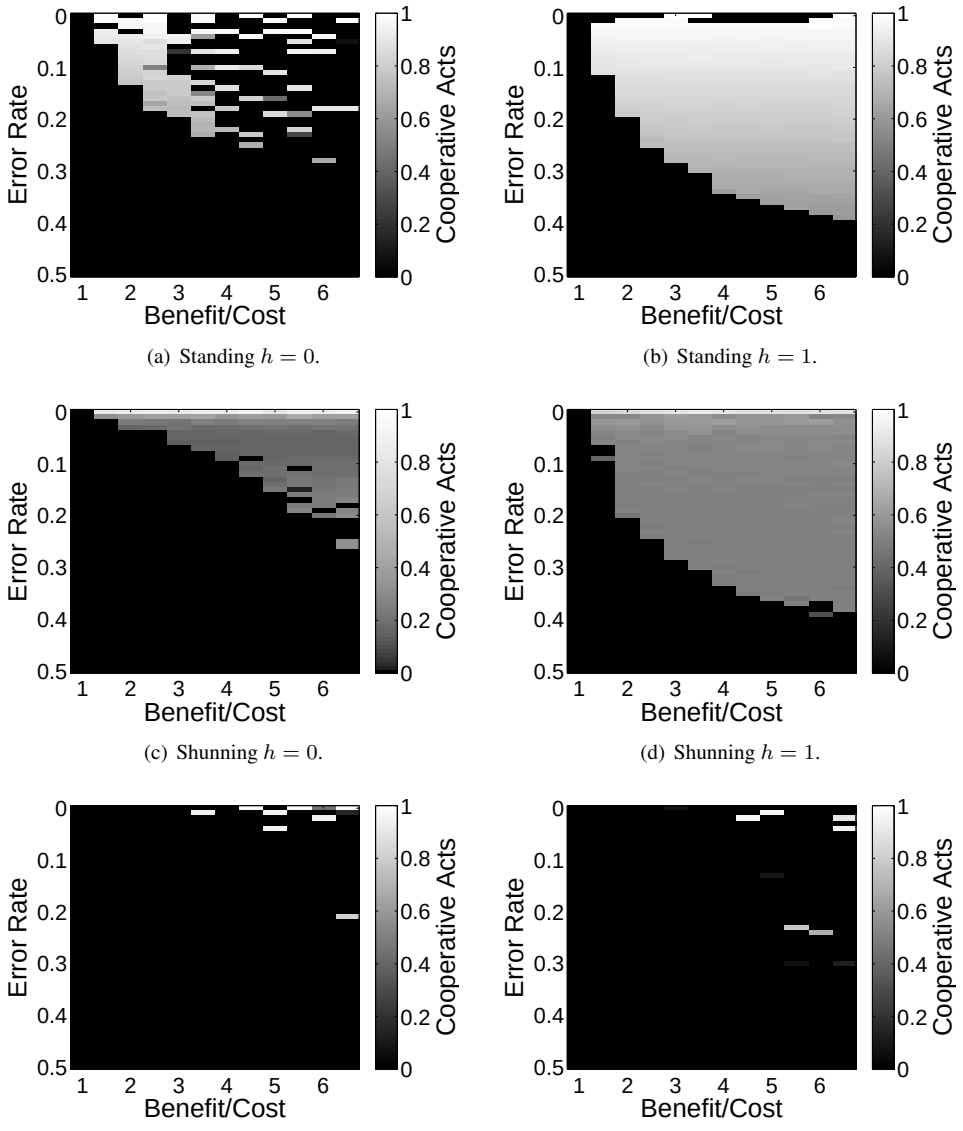
In conclusion, Figure A-2 illustrates that the results from Section 3.5 extend to additional social norms. Homophily increases cooperation for Judging and Shunning societies, which are the two norms stable in populations where agents may disagree on the reputation of an individual (Takahashi and Mashima, 2006). Furthermore, homophily increases the stability of an unstable norm in Standing.

A.1.2 Model 3: VDISC Stability

Here we test whether DISCs invade a population of VDISCs with social norms other then Judging. We use the three norms shown in Figure A-1 and described in more detail in A.1.1. As in Section 3.6, each population starts entirely comprised of VDISC agents. ALLD, ALLC, and DISC agents can enter the population through mutation. We test whether DISC agents invade.

Results

Figure A-3 extends the results shown in Section 3.6 for the Standing and Shunning social norm. The graphs in A-3 depict the fraction of DISC agents after 500 generations. Without homophily DISC agents do not invade with Standing (Figure A-3(a)), Shunning (Figure A-3(c)), or Image Scoring (Figure A-3(e)). However, when homophily is added, DISCs invade for both Standing (Figure A-3(b)) and Shunning (Figure A-3(c)). As Image Scoring does not sustain cooperation, DISCs do not invade because ALLDs invade, rather than because of VDISC stability — Figure A-3(f). For both Standing and Shunning, homophily increases a cooperative society's robustness against error in communication (see Section 3.5 and A.1.1), but consequentially selects for agents who do not employ accurate reputational information over erroneous social information.



(a) Standing $h = 0$.

(b) Standing $h = 1$.

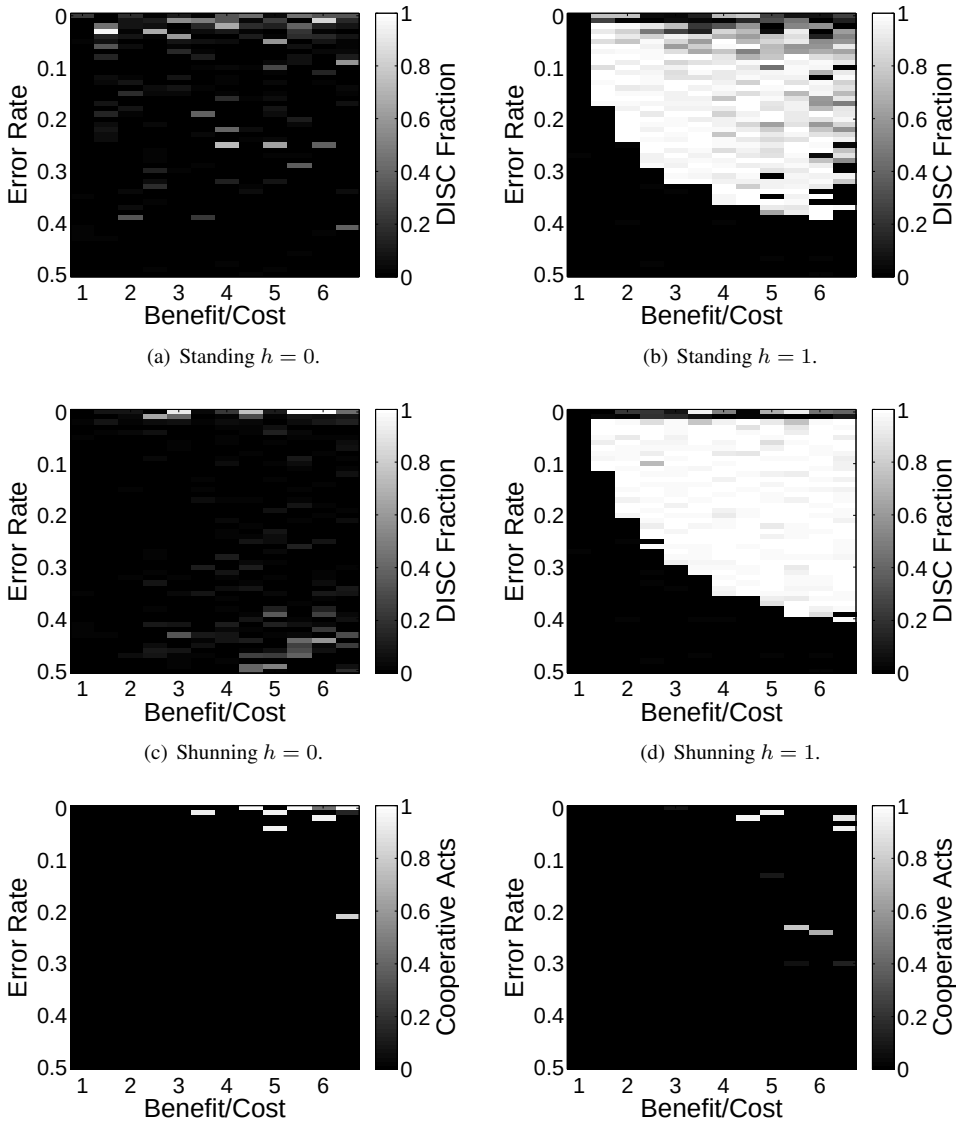
(c) Shunning $h = 0$.

(d) Shunning $h = 1$.

(e) Image Scoring $h = 0$.

(f) Image Scoring $h = 1$.

Figure A-2: Fraction of cooperative acts averaged over last 50 generations. The figures illustrate the consequences of adding homophily to three social norms.



(a) Standing $h = 0$.

(b) Standing $h = 1$.

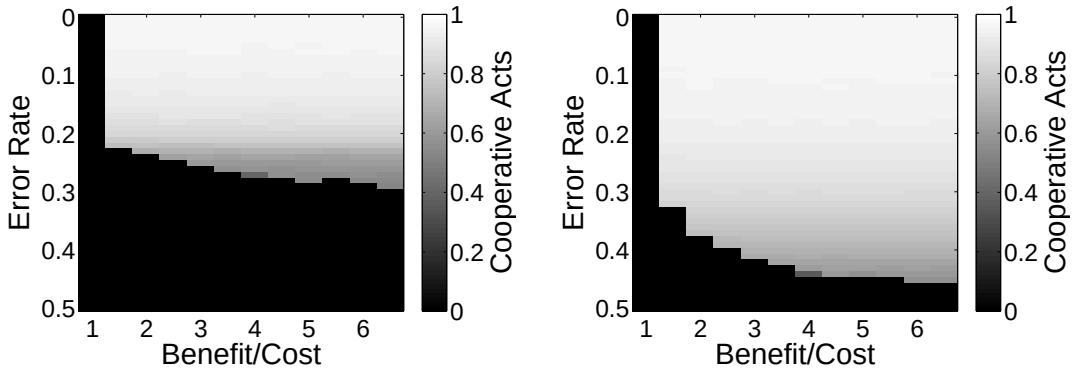
(c) Shunning $h = 0$.

(d) Shunning $h = 1$.

(e) Image Scoring $h = 0$.

(f) Image Scoring $h = 1$.

Figure A-3: Fraction of DISC agents after 500 generations. The figures illustrate whether DISC agents invade a population of VDISC using three social norms.



(a) $h = 0$ (b) $h = 1$

Figure A-4: Depicts the consequences of continuous reputations for different levels of homophily. The figures illustrate the average fraction of donation games which were cooperative over the final 50 of 500 generations. **(a)** Homophily is 0% ($h = 0$). **(b)** Homophily is 100% ($h = 1$). With continuous reputations, homophily still increases robustness against error in communication.

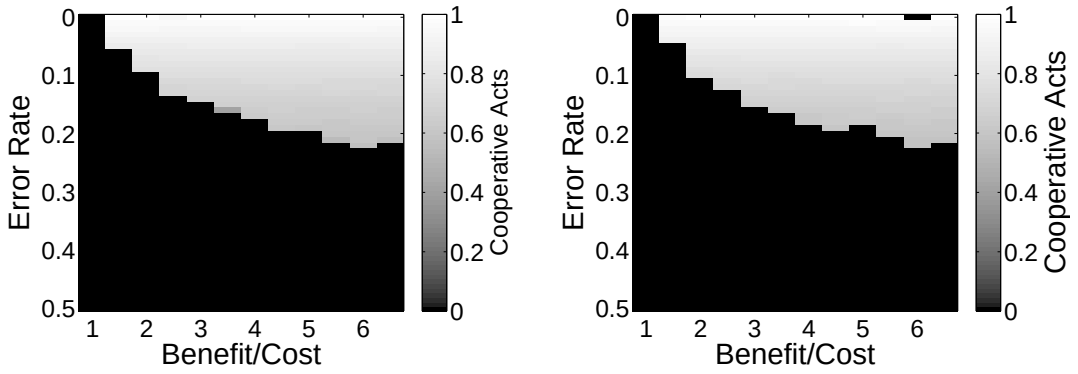
A.2 Model 2: Continuous Reputations

Up until now, an agent's reputation was based on its most recent action. If an agent had incurred a negative reputation many times in the past, it could alter its reputation with one action. Here, we test the consequences of homophily when there is some historical memory.

Each agent begins with a good reputation, $R = 1$. Depending on the judgement of the observer, the reputation of the donor moves ± 0.2. The error rate (e) affects the sign of the reputational movement. So, if the observer tells the population to update its reputation of the donor by $+0.2$, e percent of the population will update it with -0.2. When judging the recipient, the threshold between good and bad is 0.5. Thus, if a DISC donor meets a recipient with a reputation greater than 0.5, it will cooperate, otherwise it will defect. Reputation is limited in the range $[0, 1]$, so if an agent receives a reputation higher than 1, it remains at 1. For this simulation, we used the Judging norm (see Table 3.2).

Results

Figure A-4 shows the result of homophily with continuous reputations. Clearly, a homophilous society ($h = 1$) is able to sustain cooperation in the face of increased error. Figure A-4(a) illustrates that cooperation fails if $e > 0.3$, while Figure A-4(b) demonstrates that a homophilous society can maintain cooperation for an error rate near 0.45.



(a) Donor and Recipient agree

(b) Observer and Recipient agree

Figure A-5: The fraction of cooperative actions for two definitions of homophily. $h = 1$ for both, and the fraction of cooperative acts is average of the last 50 generations.

A.3 Defining Homophily

In Section 3.5 we defined value homophily as the propensity for the donor and observer to agree on the reputation of the recipient. Here we test the repercussions of redefining it. We evaluate two other definitions, namely, value homophily is the propensity for *a*) a donor and recipient to share a belief regarding another; and, *b*) an observer and recipient to share a belief regarding another.

In both cases, initially, a donor, recipient, and observer are selected at random. In the first case, in a homophilic interaction, the donor and recipient share a reputation with another. As the observer is chosen randomly, we compare the donor's belief of the observer with the recipient's belief in the observer. If the two disagree, then h is the probability that a new recipient is selected which agrees with the donor on the reputation of the observer.

In the final instantiation of homophily, an observer and recipient must share a belief in the reputation of the donor. If they do not, then h is the probability that a new observer is selected which agrees with the recipient.

Results

Figure A-5 displays the number of cooperative acts when $h = 1$ for the two definitions of homophily. Recall that Figure 3-2 represents a baseline, where all actors are randomly selected ($h = 0$). Whether homophily is defined as unanimity between donor and recipient (Figure A-5(a)) or recipient and observer (A-5(b)), homophily does not aid or hinder cooperation, relative to the baseline. This is in contrast to Figure 3-3(a), where homophily is defined as agreement between donor and observer.

A.4 Simplification

A.4.1 Simplification of DISC Stability

Here we simply the expression defining when a DISC population is stable against an ALLD invasion. We start with Equation 3.3:

$$\pi_r - \pi_d > 0 \tag{A.1}$$

Substituting Equations 3.1 and 3.2:

$$bP_r - c + \frac{w}{1-w}[bP_rG_r - c(P_rG_r + P_dG_d)] - [bP_r + \frac{w}{1-w}(bP_rG_d)] > 0 \tag{A.2}$$

Since we are testing ALLD invasion criteria, we presume that the proportion of DISCs is close to one, and the proportion of ALLDs is close to zero. In other words, $P_r \approx 1$ and $P_d \approx 0$. Replacing these:

$$b - c + \frac{w}{1-w}[bG_r - cG_r] - [b + \frac{w}{1-w}(bG_d)] > 0 \tag{A.3}$$

For simplicity, we replace $\alpha = \frac{w}{1-w}$.

$$b - c + \alpha[bG_r - cG_r] - [b + \alpha(bG_d)] > 0 \tag{A.4}$$

Dividing by α we get:

$$\frac{b-c}{\alpha} + [bG_r - cG_r] - \frac{b}{\alpha} - bG_d > 0 \tag{A.5}$$

Since $\frac{1}{\alpha} = \frac{1-w}{w}$ and $w \approx 1$, then $\frac{1}{\alpha} \approx 0$, the equation simplifies to:

$$\begin{aligned} bG_r - cG_r - bG_d &> 0 \\ G_r(b-c) - bG_d &> 0 \\ G_r &> \frac{bG_d}{b-c} \end{aligned} \tag{A.6}$$

A.4.2 Simplification of VDISC Stability

Here we simplify $\pi_v - \pi_r > 0$. Substituting Equations 3.8 and 3.9 we get:

$$b - c + \frac{w}{1-w}[b(P_rG_v + bP_vG_v) - c(P_rG_r + P_vG_v)] \\ - (b-c) - \frac{w}{1-w}[b(P_rG_r + P_vG_r) - c(P_rG_r + P_vG_v)] > 0 \quad \text{(A.7)}$$

In the first round both the VDISC and DISC agent receive the same pay-off $(b-c)$. $(b-c)-(b-c) = 0$, so they are removed. Furthermore, for simplicity, we replace $\alpha = \frac{w}{1-w}$.

$$\alpha[b(P_rG_v + bP_vG_v) - c(P_rG_r + P_vG_v)] \\ - \alpha[[b(P_rG_r + P_vG_r) - c(P_rG_r + P_vG_v)] > 0 \quad \text{(A.8)}$$

Since we are testing whether DISCs can invade, the probability of interacting with a VDISC is close to one ($P_v \approx 1$), and DISCs are rare ($P_r \approx 1$). Thus, we get:

$$\alpha[bG_v - cG_v] - \alpha[bG_r - cG_v] > 0 \quad \text{(A.9)}$$

Since $\alpha > 0$, dividing by α, simplifies the formula to:

$$bG_v - cG_v - bG_r + cG_v > 0 \quad \text{(A.10)}$$

$cG_v - cG_v = 0$. Since $b > 0$, dividing by b reduces the expression to:

$$G_v - G_r > 0 \quad \text{(A.11)}$$

Appendix B

Appendix for Chapter 8

B.1 5% and 10% Inquiry Dependency

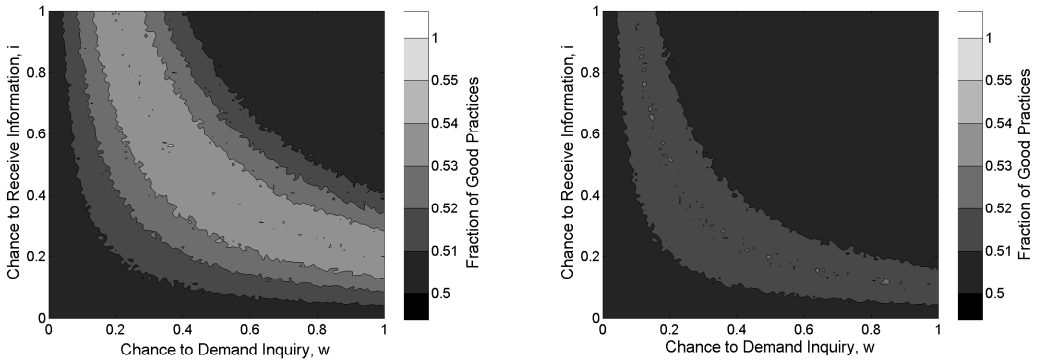
Figure 8-2 depicts patient care when review time is increased by 3% whenever a new inquiry is added to the inquiry list. Here we demonstrate how patient care is affected when each inquiry is increased by 5% and 10%. Figure B-1(a) and B-1(b) illustrate that as the interdependency between reviews rise, patient care is improved by higher levels of complicity and informational obfuscation. At 3% (Figure 8-2), if information is transparent ($i = 1$), then patients receive the best care when workers are complicit at a rate of 65-75%. However, as seen in Figure B-1(a), when inquiries are 5% dependant, then an 80% complicity rate is optimal. Further, Figure B-1(b) demonstrates that at 10%, an 85-95% rate of complicity is best.

B.2 Model 2 For 50,000 Rounds

Figure B-2 illustrates that if Model 2 is run for 50,000 rounds, then transparency and frequent whistleblowing is the best strategy for maximizing patient care. Over time, the benefits of whistleblowing increase. The ramifications of this are discussed in the main text.

B.3 5% and 10% Over 100,000 Rounds

Here we examine the consequences of decreasing the processing efficiency of formal inquiries over time. Figures B-3(a) and B-3(b) illustrate the effects of adding 5% and 10% to each inquiry's processing time when another inquiry is added to the queue. Whistleblowing at a rate of 0.3 outperforms unconditional whistleblowing ($w = 1$), until round 34,326 and 53,987 for 5% and 10% dependency, respectively. This is in contrast to round 18,199 given a 3% dependency. $w = 0.6$ outperforms unconditional whistleblowing ($w =$



(a) Patient care with a 5% dependency.

(b) Patient care with a 10% dependency.

Figure B-1: Patient care with higher dependency rates.

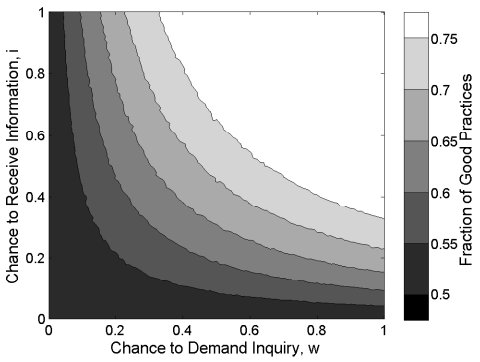
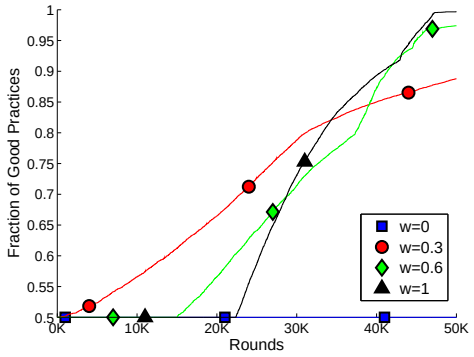
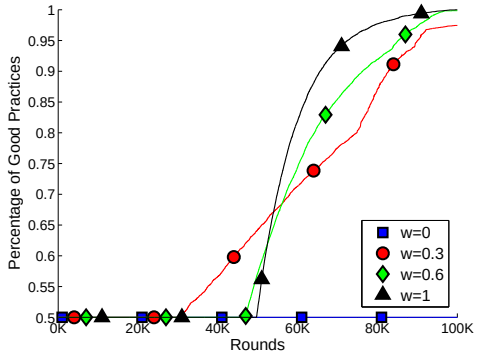


Figure B-2: Average health care after 50,000 rounds. 100 round initial processing time with a 3% dependency



(a) 5% dependency.



(b) 10% Dependency. Note that the x-axis is extended to 100,000 rounds.

Figure B-3: Patient care with higher dependency rates over 100,000 rounds.

1) until round 28,706 and 52,412 for 5% and 10% dependency. This is in contrast to round 24,800 given a 3% dependency. Clearly, decreasing processing efficiency significantly delays the utility of unconditional whistleblowing.

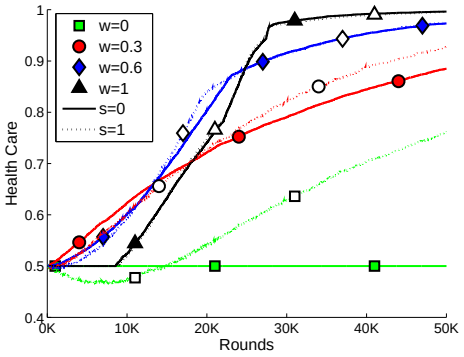
One additional result is worth noting. Unlike when the dependency rate is set to 3%, in both 5 and 10% scenarios, unconditional whistleblowing surpasses a 60% whistleblowing rate before a 30% whistleblowing rate. This demonstrates that even a simple mechanism such as processing dependencies can have complicated repercussions on the best whistleblowing strategy for patient care.

B.4 Increased Change for Harmful Soft Advice

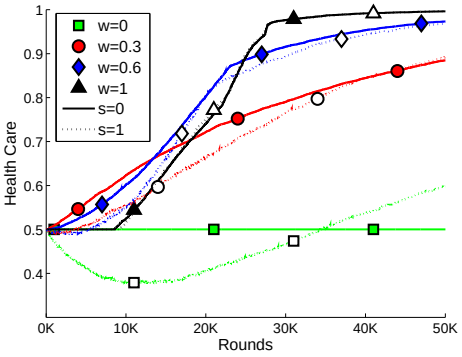
In Model 5 there is an equal probability that soft advice helps, hinders, or does not affect the questioned practice. Here we consider how the likelihood of harmful advice affects the utility of soft advice. In Figure B-4(a) and B-4(b), different probabilities for harmful advice is compared to an institution which does not permit advice.

In Figure B-4(a), when an individual softly advises, there is a 50% chance the advice will result in harmful practice (i.e. a practice with a value of -1). There is both a 25% chance the advice will result in a good practice or will not alter the practice. The result is significantly different from Model 5 where the probability of each result is equally likely. When there is a 30% or 60% chance of whistleblowing, softly advising ($s = 1$; dotted line) initially performs worse than an institution without soft advice ($s = 0$; solid line). However, eventually there is a period where softly advising surpasses an institution without soft advice.

In Figure B-4(b) there is a $2/3$ chance soft advice will prove harmful. There is a $1/6$ chance that the advice will augment or not affect the practice. Again, the results are in stark contrast to both Model 5



(a) Comparing health care when $s = 0$ (solid lines) and $s = 1$ (dotted lines). Soft advice has a $1/2$ chance to cause harm (-1), $1/4$ chance to be ignored, and a $1/4$ change to benefit the practice.



(b) Comparing health care when $s = 0$ (solid lines) and $s = 1$ (dotted lines). Soft advice has a $2/3$ chance to cause harm (-1), $1/6$ chance to be ignored, and a $1/6$ change to benefit the practice.

Figure B-4: Patient care when soft advice is increasingly harmful

and Figure B-4(a). There is no point where softly advising outperforms an institution where soft advice is prohibited. It is consistently harmful to advice peers without formal reviews.

While Model 5 illustrates that soft advice can augment formal reviews even when there is a high chance the advice will harm practice, this analysis demonstrates that there are limits beyond which soft advice may hinder patient care. Depending on the average quality of the advice, patient care may be augmented (Figure 8-5), harmed (Figure B-4(b)), or a mixture of the two (Figure B-4(a)).

Bibliography

Achtziger, A., Als-Ferrer, C., and Wagner, A. K. (2015). The impact of self-control depletion on social preferences in the ultimatum game. *Journal of Economic Psychology*, 53:1–15.

Adamic, L., Buyukkokten, O., and Adar, E. (2003). A social network caught in the web. *First Monday*, 8(6).

Al-Ubaydli, O., Houser, D., Nye, J., Paganelli, M. P., and Pan, X. S. (2013). The causal effect of market priming on trust: An experimental investigation using randomized control. *PloS one*, 8(3):e55968.

Alexander, J. M. (2009). Evolutionary game theory. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2009 edition.

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology*, 49(6):1621–1630.

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., and Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of personality and social psychology*, 68(5):804.

Alloy, L. B. and Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General*, 108(4):441–485.

Alloy, L. B. and Clements, C. M. (1992). Illusion of control: Invulnerability to negative affect and depressive symptoms after laboratory and natural stressors. *Journal of Abnormal Psychology*, 101(2):234–245.

Anderson, C., Brion, S., Moore, D. A., and Kennedy, J. A. (2012). A status-enhancement account of overconfidence. *Journal of personality and social psychology*, 103(4):718–735.

André, J.-B. and Baumard, N. (2011). The evolution of fairness in a biological market. *Evolution*, 65(5):1447–1456.

Angerer, S., Glätzle-Rützler, D., Lergetporer, P., and Sutter, M. (2014). Donations, risk attitudes and time preferences: A study on altruism in primary school children. Technical report, Ifo Institute for Economic Research at the University of Munich.

Arndt, J., Greenberg, J., Solomon, S., Pyszczynski, T., and Simon, L. (1997). Suppression, accessibility of death-related thoughts, and cultural worldview defense: Exploring the psychodynamics of terror management. *Journal of Personality and Social Psychology*, 73(1):5–18.

Asghari, F., Fotouhi, A., and Jafarian, A. (2010). Doctors' views of attitudes towards peer medical error. *Postgraduate medical journal*, 86(1012):123–6.

Attree, M. (2007). Factors influencing nurses' decisions to raise concerns about care quality. *Journal of nursing management*, 15(4):392–402.

Axelrod, R. (1997). The dissemination of culture a model with local convergence and global polarization. *Journal of conflict resolution*, 41(2):203–226.

Axelrod, R. and Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211:1390–1396.

Back, M. D., Penke, L., Schmukle, S. C., Sachse, K., Borkenau, P., and Asendorpf, J. B. (2011). Why mate choices are not as reciprocal as we assume: The role of personality, flirting and physical attractiveness. *European Journal of Personality*, 25(2):120–132.

Baghramian, M. and Nicholson, A. (2013). The puzzle of self-deception. *Philosophy Compass*, 8(11):1018–1029.

Bagnoli, C. (2012). Self-deception and agential authority. a constitutivist account. *Humanamente*, pages 99–116.

Baird, A. D. and McKay, R. T. (2008). Psychological factors in retrograde amnesia: Self-deception and a broken heart. *Neurocase*, 14(5):400–413. PMID: 18825573.

Balliet, D. and Van Lange, P. A. M. (2013). Trust, punishment, and cooperation across 18 societies: A meta-analysis. *Perspectives on Psychological Science*, 8(4):363–379.

Bandura, A. (2011). Self-deception: A paradox revisited. *Behavioral and Brain Sciences*, 34:16–17.

Barclay, P. and Stoller, B. (2014). Local competition sparks concerns for fairness in the ultimatum game. *Biology Letters*, 10(5).

Barnett, T. (1992). A preliminary investigation of the relationship between selected organizational characteristics and external whistleblowing by employees. *Journal of Business Ethics*, 11(12):949–959.

Bateson, M., Brilot, B., and Nettle, D. (2011). Anxiety: an evolutionary approach. *Canadian journal of psychiatry*, 56(12):707–715.

Baucus, M. S. and Dworkin, T. M. (1994). Wrongful firing in violation of public policy: Who gets fired and why. *Employee Responsibilities and Rights Journal*, 7(3):191–206.

Baumard, N., André, J.-B., and Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, 36(01):59–78.

Bear, A. and Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences*.

Becker, E. (1973). *The denial of death*. New York: The Free Press, New York.

Bénabou, R. (2013). Groupthink: Collective delusions in organizations and markets. *The Review of Economic Studies*, 80(2):429–462.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142.

Bermúdez, J. L. (2000). Self-deception, intentions, and contradictory beliefs. *Analysis*, 60(268):309–319.

Bisiach, E., Rusconi, M. L., and Vallar, G. (1991). Remission of somatoparaphrenic delusion through vestibular stimulation. *Neuropsychologia*, 29(10):1029 – 1031.

Black, L. M. (2011). Original Research: Tragedy into Policy: A Quantitative Study of Nurses' Attitudes Toward Patient Advocacy Activities. *AJN, American Journal of Nursing:*, 111(6):26–35.

Boero, R., Bravo, G., Castellani, M., and Squazzoni, F. (2009). Reputational cues in repeated trust games. *The Journal of Socio-Economics*, 38(6):871–877.

Bolsin, S., Faunce, T., and Oakley, J. (2005). Practical virtue ethics: healthcare whistleblowing and portable digital technology. *Journal of medical ethics*, 31(10):612–8.

Bond, CharlesF., J. and Robinson, M. (1988a). The evolution of deception. *Journal of Nonverbal Behavior*, 12(4):295–307.

Bond, CharlesF., J. and Robinson, M. (1988b). The evolution of deception. *Journal of Nonverbal Behavior*, 12(4):295–307.

Bone, J. E. and Raihani, N. J. (2015). Human punishment is motivated by both a desire for revenge and a desire for equality. *Evolution and Human Behavior*, 36(4):323–330.

Bortolotti, L. and Mameli, M. (2012). Self-deception, delusion and the boundaries of folk psychology. *Humanamente*, 20:203–221.

Boudry, M., Vlerick, M., and McKay, R. (2015). Can evolution get us off the hook? evaluating the ecological defence of human rationality. *Consciousness and Cognition*, 33:524 – 535.

Boyd, R. and Richerson, P. J. (1985). Culture and the evolutionary process.

Boyd, R. and Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology*, 13(3):171 – 195.

Bravo, G. and Tamburino, L. (2008). The evolution of trust in non-simultaneous exchange situations. *Rationality and Society*, 20(1):85–113.

Brickman, P., Coates, D., and Janoff-Bulman, R. (1978). Lottery winners and accident victims: Is happiness relative? *Journal of Personality and Social Psychology*, 36(8):917–927.

Brown, J. D. (2014). Self-esteem and self-evaluation: Feeling is believing. In *Psychological Perspectives on the Self, Volume 4: The Self in Social Perspective*, volume 4, pages 27–58.

Bryson, J. J. (2009). Representations underlying social learning and cultural evolution. *Interaction Studies*, 10(1):77–100.

Bryson, J. J., Lowe, W., Bilovich, A., and Čače, I. (in prep). Factors determining the extent of a species' reliance on socially-acquired behavior. in prep.

Buechel, E. C., Zhang, J., Morewedge, C. K., and Vosgerau, J. (2014). More intense experiences, less intense forecasts: Why people overweight probability specifications in affective forecasts. *Journal of Personality and Social Psychology*, 106(1):20–36.

Buehler, R. and McFarland, C. (2001). Intensity bias in affective forecasting: The role of temporal focus. *Personality and Social Psychology Bulletin*, 27(11):1480–1493.

Burks, S. V., Carpenter, J. P., and Verhoogen, E. (2003). Playing both roles in the trust game. 51(2):195–216.

Buss, D. M. (2007). The evolution of human mating. *Acta Psychologica Sinica*, 39(3):502–512.

Buss, D. M. and Kenrick, D. T. (1998). Evolutionary social psychology.

Butler, P. V. (2000). Reverse othello syndrome subsequent to traumatic brain injury. *Psychiatry*, 63(1):85–92.

Byrne, D. (1961). Interpersonal attraction and attitude similarity. *The Journal of Abnormal and Social Psychology*, 62(3):713.

Callahan, E. S. and Dworkin, T. M. (1992). Do Good and Get Rich: Financial Incentives for Whistleblowing and the False Claims Act. *Villanova Law Review*, 37:273.

Chance, Z., Gino, F., Norton, M. I., and Ariely, D. (2015). The slow decay and quick revival of self-deception. *Frontiers in Psychology*, 6(1–6).

Chance, Z. and Norton, M. I. (2015). The what and why of self-deception. *Current Opinion in Psychology*, 6:104 – 107.

Chance, Z., Norton, M. I., Gino, F., and Ariely, D. (2011). Temporal view of the costs and benefits of self-deception. *Proceedings of the National Academy of Sciences*, 108(Supplement 3):15655–15659.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

Chen, X., Szolnoki, A., and Perc, M. (2014). Probabilistic sharing solves the problem of costly punishment. *New Journal of Physics*, 16(8):083016.

Chiang, Y.-S. (2010). Self-interested partner selection can lead to the emergence of fairness. *Evolution and Human Behavior*, 31(4):265 – 270.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, 78(2):67–90.

Clark, A. E., Diener, E., Georgellis, Y., and Lucas, R. E. (2008). Lags and leads in life satisfaction: A test of the baseline hypothesis*. *The Economic Journal*, 118(529):F222–F243.

Cloak, F.T., J. (1975). Is a cultural ethology possible? *Human Ecology*, 3(3):161–182.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? studies with the wason selection task. *Cognition*, 31(3):187 – 276.

Cosmides, L. and Tooby, J. (1997). Evolutionary psychology: A primer. *Retrieved January*, 1:2004.

Cosmides, L. and Tooby, J. (2013). Evolutionary psychology: New perspectives on cognition and motivation. *Annual Review of Psychology*, 64(1):201–229. PMID: 23282055.

Cosmides, L., Tooby, J., and Barkow, J. H. (1992). Introduction: Evolutionary psychology and conceptual integration. In Barkow, J. H., Cosmides, L., and Tooby, J., editors, *The Adapted Mind*, pages 163–228. Oxford University Press.

Creswell, J. D., Lam, S., Stanton, A. L., Taylor, S. E., Bower, J. E., and Sherman, D. K. (2007). Does self-affirmation, cognitive processing, or discovery of meaning explain cancer-related health benefits of expressive writing? *Personality and Social Psychology Bulletin*, 33(2):238–250.

Creswell, J. D., Welch, W. T., Taylor, S. E., Sherman, D. K., Gruenewald, T. L., and Mann, T. (2005). Affirmation of personal values buffers neuroendocrine and psychological stress responses. *Psychological Science*, 16(11):846–851.

Cummins, D. D. (1999). Cheater detection is modified by social rank: The impact of dominance on the evolution of cognitive functions. *Evolution and Human Behavior*, 20(4):229 – 248.

Cummins, R. and Nistico, H. (2002). Maintaining life satisfaction: The role of positive cognitive bias. *Journal of Happiness Studies*, 3(1):37–69.

Curry, O. and Dunbar, R. I. (2013). Do birds of a feather flock together? *Human Nature*, pages 1–12.

Czerlinski, J., Gigerenzer, G., and Goldstein, D. G. (1999). How good are simple heuristics? In Gigerenzer, G. and Todd, P. M., editors, *Simple heuristics that make us smart. Evolution and cognition.*, pages 97–118. Oxford University Press.

Dalziel, J. R. and Job, R. (1997). Motor vehicle accidents, fatigue and optimism bias in taxi drivers. *Accident Analysis & Prevention*, 29(4):489–494.

Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892.

Darwin, C. (1872). *On the Origin of Species, 6th Edition*.

Davidson, D. (1987). 3. deception and division. *The multiple self*, page 79.

Dawkins, R. (1976). *The Selfish Gene*. Oxford University Press.

Dawkins, R. (1982). *The Extended Phenotype: The Gene As the Unit of Selection*. W.H. Freeman & Company.

Debove, S., André, J.-B., and Baumard, N. (2015). Partner choice creates fairness in humans. *Proceedings of the Royal Society of London B: Biological Sciences*, 282(1808).

DeJoy, D. M. (1989). The optimism bias and traffic accident risk perception. *Accident Analysis & Prevention*, 21(4):333–340.

Dekker, S. W. A. and Hugh, T. B. (2014). A just culture after mid staffordshire. *BMJ quality & safety*, 23(5):356–358.

Delgado, M. R., Frank, R. H., and Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature neuroscience*, 8(11):1611–1618.

Demos, R. (1960). Lying to oneself. *The Journal of Philosophy*, 57(18):588–595.

Dennett, D. C. (1987). *The Intentional Stance*. The MIT Press, Massachusetts.

Dennett, D. C. (1995). *Darwin's Dangerous Idea*. Penguin.

DesRoches, C. M., Rao, S. R., Fromson, J. A., Birnbaum, R. J., Iezzoni, L., Vogeli, C., and Campbell, E. G. (2010). Physicians' perceptions, preparedness for reporting, and experiences related to impaired and incompetent colleagues. *JAMA*, 304(2):187–93.

Dessalles, J.-L. (2007). *Why we talk: the evolutionary origins of language*. Oxford University Press.

Deweese-Boyd, I. (2012). Self-deception. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2012 edition.

Dieckmann, U. and Doebeli, M. (1999). On the origin of species by sympatric speciation. *Nature*, 400(6742):354–357.

Dillard, A., Fagerlin, A., Cin, S. D., Zikmund-Fisher, B. J., and Ubel, P. A. (2010). Narratives that address affective forecasting errors reduce perceived barriers to colorectal cancer screening. *Social Science & Medicine*, 71(1):45–52.

dos Santos, M. (2014). The evolution of anti-social rewarding and its countermeasures in public goods games. *Proceedings of the Royal Society of London B: Biological Sciences*, 282(1798).

Dufner, M., Denissen, J., Sedikides, C., Van Zalk, M., Meeus, W. H. J., and van Aken, M. (2013). Are actual and perceived intellectual self-enhancers evaluated differently by social perceivers? *European Journal of Personality*, 27(6):621–633.

Dufner, M., Denissen, J. J. A., van Zalk, M., Matthes, B., Meeus, W. H. J., van Aken, M. A. G., and Sedikides, C. (2012). Positive intelligence illusions: On the relation between intellectual self-enhancement and psychological adjustment. *Journal of Personality*, 80(3):537–572.

Dunn, E. W., Brackett, M. A., Ashton-James, C., Schneiderman, E., and Salovey, P. (2007). On emotionally intelligent time travel: Individual differences in affective forecasting ability. *Personality and Social Psychology Bulletin*, 33(1):85–93.

Dunning, D. (2011). Get thee to a laboratory. *Behavioral and Brain Sciences*, 34:18–19.

Easton, J. A., Schipper, L. D., and Shackelford, T. K. (2007). Morbid jealousy from an evolutionary psychological perspective. *Evolution and Human Behavior*, 28(6):399 – 402.

Eastwick, P. W., Finkel, E. J., Krishnamurti, T., and Loewenstein, G. (2008). Mispredicting distress following romantic breakup: Revealing the time course of the affective forecasting error. *Journal of Experimental Social Psychology*, 44(3):800–807.

Eaton, T. V. and Akers, M. D. (2007). Whistleblowing and Good Governance. *The CPA Journal*, 77(6):66–71.

Ekman, P. and O'Sullivan, M. (1991). Who can catch a liar? *American Psychologist*, 46(9):913–920.

Ellwardt, L., Labianca, G. J., and Wittek, R. (2012). Who are the objects of positive and negative gossip at work? A social network perspective on workplace gossip. *Social Networks*, 34(2):193–205.

Emanuel, A. S., Updegraff, J. A., Kalmbach, D. A., and Ciesla, J. A. (2010). The role of mindfulness facets in affective forecasting. *Personality and Individual Differences*, 49(7):815 – 818.

Emler, N. (1990). A social psychology of reputation. *European Review of Social Psychology*, 1(1):171–193.

Engle-Warnick, J. and Slonim, R. L. (2004). The evolution of strategies in a repeated trust game. *Journal of Economic Behavior & Organization*, 55(4):553–573.

English, S., Browning, L. E., and Raihani, N. J. (2015). Developmental plasticity and social specialization in cooperative societies. *Animal Behaviour*, 106:37–42.

Enquist, M. and Leimar, O. (1993). The evolution of cooperation in mobile organisms. *Animal Behaviour*, 45:747–757.

Espín, A. M., Exadaktylos, F., Herrmann, B., and Brañas-Garza, P. (2015). Short-and long-run goals in ultimatum bargaining: impatience predicts spite-based behavior. *Frontiers in behavioral neuroscience*, 9.

Fallis, D. (2010). Lying and deception. *Philosophers' Imprint*, 10(11).

Faunce, T. (2004). Developing and teaching the virtue-ethics foundations of healthcare whistle blowing. *Monash Bioethics Review*, 23(4):41–55.

Fawcett, T. W., Fallenstein, B., Higginson, A. D., Houston, A. I., Mallpress, D. E., Trimmer, P. C., and McNamara, J. M. (2014). The evolution of decision rules in complex environments. *Trends in Cognitive Sciences*, 18:153–161.

Fehr, E., Fischbacher, U., and Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, 13(1):1–25.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):pp. 817–868.

Fernández-Juricic, E. and Kacelnik, A. (2004). Information transfer and gain in flocks: The effects of quality and quantity of social information at different neighbour distances. *Behavioral Ecology and Sociobiology*, 55(5):502–511.

Fetchenhauer, D. and Dunning, D. (2010). Why so cynical? Asymmetric feedback underlies misguided skepticism regarding the trustworthiness of others. *Psychological Science*, 21(2):189–193.

Feys, M. and Anseel, F. (2015). When idols look into the future: Fair treatment modulates the affective forecasting error in talent show candidates. *British Journal of Social Psychology*, 54(1):19–36.

Fingarette, H. (1969). *Self-deception*. University of California Pr.

Fiore, A. T. and Donath, J. S. (2005). Homophily in online dating: when do you like someone like yourself? In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, pages 1371–1374. ACM.

Fischbacher, U., Hertwig, R., and Bruhin, A. (2013). How to model heterogeneity in costly punishment: Insights from responders' response times. *Journal of Behavioral Decision Making*, 26(5):462–476.

Fishman, M. A. (2003). Indirect reciprocity among imperfect individuals. *Journal of Theoretical Biology*, 225(3):285 – 292.

Flower, T. P., Gribble, M., and Ridley, A. R. (2014). Deception by flexible alarm mimicry in an african bird. *Science*, 344(6183):513–516.

Francis, R. (2015). Freedom to Speak Up. An independent review into creating an open and honest reporting culture in the NHS.

Frankish, K. (2009). Adaptive misbelief or judicious pragmatic acceptance? *Behavioral and Brain Sciences*, 32:520–521.

Frederick, S. and Loewenstein, G. (1999). Hedonic adaptation. In Kahneman, D., Diener, E., and Schwarz, N., editors, *Well-being: The foundations of hedonic psychology.*, pages 302–329. Russell Sage Foundation.

Fredrickson, B. L., Kahneman, D., et al. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of personality and social psychology*, 65:45–55.

Freestone, L., Bolsin, S. N., Colson, M., Patrick, A., and Creati, B. (2006). Voluntary incident reporting by anaesthetic trainees in an Australian hospital. *International Journal for Quality in Health Care*, 18(6):452–457.

Freud, A. (1936). *The ego and the mechanisms of defence*. Karnac Books.

Fridland, E. (2011). Reviewing the logic of self-deception. *Behavioral and Brain Sciences*, 34:22–23.

Fu, F., Nowak, M. A., Christakis, N. A., and Fowler, J. H. (2012). The evolution of homophily. *Scientific reports*, 2:845.

Fudenberg, D., David, R., and Dreber, A. (2012). Slow to anger and fast to forgive: Cooperation in an uncertain world. *The American Economic Review*, 102(2):720–749.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, 102(4):652.

Galperin, A. and Haselton, M. G. (2012). The evolution of cognitive bias. In Forgas, J., Fiedler, K., and Sedikedes, C., editors, *Social thinking and interpersonal behavior*, pages 45–64.

Gaskett, A. C. (2011). Orchid pollination by sexual deception: pollinator perspectives. *Biological Reviews*, 86(1):33–75.

Gendler, T. S. (2008a). Alief and belief. *The Journal of Philosophy*, 105(10):634–663.

Gendler, T. S. (2008b). Alief in action (and reaction). *Mind & Language*, 23(5):552–585.

Ghislandi, P. G., Albo, M. J., Tuni, C., and Bilde, T. (2014). Evolution of deceit by worthless donations in a nuptial gift-giving spider. *Current Zoology*, 60(1):43–51.

Giardini, F. and Conte, R. (2012). Gossip for social control in natural and artificial societies. *SIMULATION*, 88(1):18–32.

Gigerenzer, G. and Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1):107–143.

Gigerenzer, G. and Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, 103(4):650–659.

Gilbert, D. T., Driver-Linn, E., and Wilson, T. D. (2002). The trouble with Vronsky: Impact bias in the forecasting of future affective states. In Salovey, L. F. B. . P., editor, *The wisdom in feeling: Psychological processes in emotional intelligence*, pages 114–143. Guildford Press.

Gilbert, D. T. and Ebert, J. E. (2002). Decisions and revisions: the affective forecasting of changeable outcomes. *Journal of personality and social psychology*, 82(4):503.

Gilbert, D. T., Killingsworth, M. A., Eyre, R. N., and Wilson, T. D. (2009). The surprising power of neighborly advice. *Science*, 323(5921):1617–1619.

Gilbert, D. T., Lieberman, M. D., Morewedge, C. K., and Wilson, T. D. (2004). The peculiar longevity of things not so bad. *Psychological Science*, 15(1):14–19.

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., Wheatley, T. P., et al. (1998). Immune neglect: A source of durability bias in affective forecasting. *Journal of personality and social psychology*, 75:617–638.

Gilbert, D. T. and Wilson, T. D. (2009). Why the brain talks to itself: sources of error in emotional prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1335–1341.

Gilmore, D. (1978). Varieties of gossip in a spanish rural community. *Ethnology*, 17(1):pp. 89–99.

Gintis, H., Bowles, S., Boyd, R., and Fehr, E. (2005). Moral sentiments and material interests: Origins, evidence, and consequences. In Gintis, H., Bowles, S., Boyd, R., and Fehr, E., editors, *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, chapter 1, pages 3–39. MIT, Cambridge, MA.

Goldberg, D. and Deb, K. (1991). A comparative analysis of selection schemes used in genetic algorithms. *Urbana*, 51:61801–2996.

Goldstein, D. G. and Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In Gigerenzer, G. and Todd, P. M., editors, *Simple heuristics that make us smart*, pages 37–58. Oxford University Press.

Goldstein, D. G. and Gigerenzer, G. (2002). Models of ecological rationality: the recognition heuristic. *Psychological Review*, 109(1):75–90.

Graham, L. and Oswald, A. J. (2010). Hedonic capital, adaptation and resilience. *Journal of Economic Behavior & Organization*, 76(2):372 – 384.

Greenberg, J., Pyszczynski, T., and Solomon, S. (1986). The causes and consequences of a need for self-esteem: A terror management theory. In *Public self and private self*, pages 189–212. Springer.

Greenberg, J., Solomon, S., and Pyszczynski, T. (1997). *Terror management theory of self-esteem and cultural worldviews: Empirical assessments and conceptual refinements.* Academic Press.

Greitemeyer, T. (2009). The effect of anticipated affect on persistence and performance. *Personality and Social Psychology Bulletin*, 35(2):172–186.

Greitemeyer, T., Lebek, S., Frey, D., and Traut-Mattausch, E. (2011). Why people try to actively change unchangeable situations: The role of anticipated affect. *Current Psychology*, 30(3):284–298.

Grimm, V. and Mengel, F. (2011). Let me sleep on it: Delay reduces rejection rates in ultimatum games. *Economics Letters*, 111(2):113 – 115.

Gruber, R., Laviolette, R., Deluca, P., Monson, E., Cornish, K., and Carrier, J. (2010). Short sleep duration is associated with poor performance on {IQ} measures in healthy school-age children. *Sleep Medicine*, 11(3):289 – 294.

Guarnaccia, C. A. (2012). *Affective forceasting: The effects of immune neglect and surrogation*. PhD thesis, University of North Texas.

Güth, W. and Kliemt, H. (2000). Evolutionarily stable co-operative commitments. *Theory and Decision*, 49(3):197–222.

Harnad, S. (2011). Deceiving ourselves about self-deception. *Behavioral and Brain Sciences*, 34:25–26.

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., and Moore, T. (2009). Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition*, 27:733–763.

Haselton, M. G. and Buss, D. M. (2000). Error management theory: a new perspective on biases in cross-sex mind reading. *Journal of personality and social psychology*, 78(1):81.

Haselton, M. G. and Buss, D. M. (2009). Error management theory and the evolution of misbeliefs. *Behavioral and Brain Sciences*, 32:522–523.

Haselton, M. G. and Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and social psychology Review*, 10(1):47–66.

Haselton, M. G., Nettle, D., and Murray, D. (2015). The evolution of cognitive bias. In Buss, M., editor, *The Evolutionary Psychology Handbook, 2nd Edition*. Springer.

Hauser, O. P., Nowak, M. A., and Rand, D. G. (2014). Punishment does not promote cooperation under exploration dynamics when anti-social punishment is possible. *Journal of Theoretical Biology*, 360:163 – 171.

Henrich, J. and Boyd, R. (1998). The evolution of conformist transmission and the emergence of between-group differences. *Evolution and human behavior*, 19(4):215–241.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. (2001). In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review*, 91(2):73–78.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., and Tracer, D. (2005). economic man in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28:795–815.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., and Ziker, J. (2010). Markets, religion, community size, and the evolution of fairness and punishment. *Science*, 327(5972):1480–1484.

Henrich, J. and Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and human behavior*, 22(3):165–196.

Herrmann, B., Thöni, C., and Gächter, S. (2008a). Antisocial punishment across societies. *Science*, 319(5868):1362–1367.

Herrmann, B., Thöni, C., and Gächter, S. (2008b). Antisocial punishment across societies. *Science*, 319(5868):1362–1367.

Herrmann, E., Keupp, S., Hare, B., Vaish, A., and Tomasello, M. (2012). Direct and indirect reputation formation in nonhuman great apes and human children. *Journal of Comparative Psychology*, (FIXME).

Higginson, A. D., Fawcett, T. W., and Houston, A. I. (2015). Evolution of a flexible rule for foraging that copes with environmental variation. *Current Zoology*, 61:303–312.

Hilbe, C., Nowak, M. A., and Sigmund, K. (2013). Evolution of extortion in iterated prisoners dilemma games. *Proceedings of the National Academy of Sciences*, 110(17):6913–6918.

Hiraishi, K., Shikishima, C., Yamagata, S., and Ando, J. (2015). Heritability of decisions and outcomes of public goods games. *Frontiers in Psychology*, 6:373.

Hoerger, M., Chapman, B. P., Epstein, R. M., and Duberstein, P. R. (2012). Emotional intelligence: A theoretical framework for individual differences in affective forecasting. *Emotion*, 12(4):716–725.

Hoerger, M., Quirk, S. W., Lucas, R. E., and Carr, T. H. (2009). Immune neglect in affective forecasting. *Journal of Research in Personality*, 43(1):91–94.

Hoerger, M., Quirk, S. W., Lucas, R. E., and Carr, T. H. (2010). Cognitive determinants of affective forecasting errors. *Judgment and decision making*, 5(5):365.

Hopfensitz, A. and Reuben, E. (2009). The importance of emotions for the effectiveness of social punishment. *The Economic Journal*, 119(540):1534–1559.

Hsee, C. and Zhang, J. (2004). Distinction bias: Misprediction and mischoice due to joint evaluation. *Journal of personality and social psychology*, 86(5).

Huck, S. and Oechssler, J. (1999). The indirect evolutionary approach to explaining fair allocations. *Games and Economic Behavior*, 28(1):13–24.

Ingram, G. P. D. and Bering, J. M. (2010). Children's tattling: The reporting of everyday norm violations in preschool settings. *Child Development*, 81(3):945–957.

Jackson, D., Peters, K., Andrew, S., Edenborough, M., Halcomb, E., Luck, L., Salamonson, Y., Weaver, R., and Wilkes, L. (2014). Trial and retribution: A qualitative study of whistleblowing and workplace relationships in nursing. *Contemporary Nurse*.

Janis, I. L. (1971). Groupthink. *Psychology Today*, 5(6):43–46.

Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes.* Houghton Mifflin.

Janz, N. K., Lakhani, I., Vijan, S., Hawley, S. T., Chung, L. K., and Katz, S. J. (2007). Determinants of colorectal cancer screening use, attempts, and non-use. *Preventive Medicine*, 44(5):452 – 458.

Jaspers, K., Hoenig, J., and Hamilton, M. W. (1963). *General psychopathology*, volume 2. JHU Press.

Jin, X.-H., Ren, Z.-X., Xu, S.-Z., Wang, H., Li, D.-Z., and Li, Z.-Y. (2014). The evolution of floral deception in epipactis veratrifolia (orchidaceae): from indirect defense to pollination. *BMC Plant Biology*, 14(1):63.

Johnson, D. D. (2009). The error of god: Error management theory, religion, and the evolution of cooperation. In *Games, groups, and the global good*, pages 169–180. Springer.

Johnson, D. D., Blumstein, D. T., Fowler, J. H., and Haselton, M. G. (2013). The evolution of error: error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology & Evolution*.

Johnson, D. D. and Fowler, J. H. (2011). The evolution of overconfidence. *Nature*, 477(7364):317–320.

Johnson, N. D. and Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5):865–889.

Jonas, E., Schulz-Hardt, S., Frey, D., and Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *Journal of personality and social psychology*, 80(4):557.

Jones, A. and Kelly, D. (2013). When care is needed: the role of whistleblowing in promoting best standards from an individual and organizational perspectivenull. *Quality in Ageing and Older Adults*, 14(3):180–191.

Jones, A. and Kelly, D. (2014a). Deafening silence? Time to reconsider whether organisations are silent or deaf when things go wrong. *BMJ quality & safety*, 23(9):709–13.

Jones, A. and Kelly, D. (2014b). Whistle-blowing and workplace culture in older peoples' care: qualitative insights from the healthcare and social care workforce. *Sociology of health & illness*, 36(7):986–1002.

Jusup, M., Matsuo, T., and Iwasa, Y. (2014). Barriers to cooperation aid ideological rigidity and threaten societal collapse. *PLoS Comput Biol*, 10(5):e1003618.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., Krueger, A. B., Schkade, D., Schwarz, N., and Stone, A. A. (2006). Would you be happier if you were richer? a focusing illusion. *Science*, 312(5782):1908–1910.

Kahneman, D. and Sugden, R. (2005). Experienced utility as a standard of policy evaluation. *Environmental and resource economics*, 32(1):161–181.

Kahneman, D., Wakker, P. P., and Sarin, R. (1997). Back to bentham? explorations of experienced utility. *The Quarterly Journal of Economics*, 112(2):375–406.

Kandel, D. B. (1978). Homophily, selection, and socialization in adolescent friendships. *American journal of Sociology*, pages 427–436.

Karagonlar, G. and Kuhlman, D. M. (2013). The role of social value orientation in response to an unfair offer in the ultimatum game. *Organizational Behavior and Human Decision Processes*, 120(2):228 – 239. Social Dilemmas.

Kennedy, J. A., Anderson, C., and Moore, D. A. (2013). When overconfidence is revealed to others: Testing the status-enhancement theory of overconfidence. *Organizational Behavior and Human Decision Processes*, 122(2):266 – 279.

King, A. J. and Cowlishaw, G. (2007). When to use social information: the advantage of large group size in individual decision making. *Biology Letters*, 3(2):137–139.

King, G. (1999). The implications of an organization's structure on whistleblowing. *Journal of Business Ethics*, 20(4):315–326.

Kingston, M., Evans, S., Smith, B., and Berry, J. (4). Attitudes of doctors and nurses towards incident reporting: a qualitative analysis. *Medical Journal of Australia1*, 181(1):36–39.

Kipp, D. (1980). On self-deception. *The Philosophical Quarterly*, 30(121):pp. 305–317.

Klein, N. and Epley, N. (2015). Group discussion improves lie detection. *Proceedings of the National Academy of Sciences*, 112(24):7460–7465.

Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., and Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, 435(7042):673–676.

Krasnow, M., Delton, A., Cosmides, L., and Tooby, J. (2015). Group cooperation without group selection: Modest punishment can recruit much cooperation. *PLoS ONE*, 10(4):e0124561.

Kurzban, R. (2011). Two problems with 'self-deception': No 'self' and no 'deception'. *Behavioral and Brain Sciences*, 34:32–33.

Kurzban, R. (2012). *Why everyone (else) is a hypocrite: Evolution and the modular mind*. Princeton University Press.

Kurzban, R. and Athena Aktipis, C. (2007). Modularity and the social mind: Are psychologists too self-ish? *Personality and Social Psychology Review*, 11(2):131–149.

Kurzban, R. and Leary, M. R. (2001). Evolutionary origins of stigmatization: the functions of social exclusion. *Psychological bulletin*, 127(2):187.

Kushlev, K. and Dunn, E. W. (2012). Affective forecasting: Knowing how we will feel in the future. In Vazire, S. and Wilson, T. D., editors, *Handbook of Self-Knowledge*, pages 277–292. Guilford Press.

Kwong, J. Y., Wong, K. F. E., and Tang, S. K. (2013). Comparing predicted and actual affective responses to process versus outcome: An emotion-as-feedback perspective. *Cognition*, 129(1):42–50.

Lacey, H. P., Smith, D. M., and Ubel, P. A. (2006). Hope i die before i get old: Mispredicting happiness across the adult lifespan. *Journal of Happiness Studies*, 7(2):167–182.

Lamba, S. and Mace, R. (2012). The evolution of fairness: explaining variation in bargaining behaviour. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1750).

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology*, 32(2):311–328.

Lazar, A. (1999). Deceiving oneself or self-deceived? on the formation of beliefs 'under the influence'. *Mind*, 108(430):265–290.

Lee, M. D., Loughlin, N., and Lundberg, I. B. (2002). Applying one reason decision-making: the prioritisation of literature searches. *Australian Journal of Psychology*, 54(3):137–143.

Levine, L. J., Lench, H. C., Kaplan, R. L., and Safer, M. A. (2012). Accuracy and artifact: Reexamining the intensity bias in affective forecasting. *Journal of Personality and Social Psychology*, 103(5):584–605.

Levine, T. R. (2015). New and improved accuracy findings in deception detection research. *Current Opinion in Psychology*, 6:1 – 5.

Loewenstein, G. (2007). Affect regulation and affective forecasting. *Handbook of emotion regulation*, pages 180–203.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., and Welch, N. (2001). Risk as feelings. *Psychological bulletin*, 127(2):267.

Lorenz, J., Rauhut, H., Schweitzer, F., and Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020–9025.

Magurran, A. and Higham, A. (1988). Information transfer across fish shoals under predator threat. *Ethology*, 78(2):153–158.

Mahon, J. E. (2015). The definition of lying and deception. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2015 edition.

Manapat, M. L., Nowak, M. A., and Rand, D. G. (2012). Information, irrationality, and the evolution of trust. *Journal of Economic Behavior & Organization*.

Manapat, M. L. and Rand, D. G. (2012). Delayed and inconsistent information and the evolution of trust. *Dynamic Games and Applications*, 2(4):401–410.

Mannion, R. and Davies, H. T. (2015). Cultures of silence and cultures of voice: The role of whistleblowing in healthcare organisations. *International Journal of Health Policy and Management*, 4(8):503–505.

Mannion, R., Davies, H. T. O., and Marshall, M. N. (2005). Cultural characteristics of high and low performing hospitalsnull. *Journal of Health Organization and Management*, 19(6):431–439.

Marks, G. and Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin*, 102(1):72–90.

Marlowe, F. W., Berbesque, J. C., Barrett, C., Bolyanatz, A., Gurven, M., and Tracer, D. (2010). The 'spiteful' origins of human cooperation. *Proceedings of the Royal Society of London B: Biological Sciences*.

Marroquín, B., Nolen-Hoeksema, S., and Miranda, R. (2013). Escaping the future: Affective forecasting in escapist fantasy and attempted suicide. *Journal of Social and Clinical Psychology*, 32(4):446–463.

Marshall, J., Trimmer, P. C., and Houston, A. I. (2013a). Unbiased individuals use valuable information when making decisions: a reply to johnson and fowler. *Trends in ecology & evolution*, 28(8):444.

Marshall, J. A. (2009). The donation game with roles played between relatives. *Journal of Theoretical Biology*, 260(3):386–391.

Marshall, J. A. (2011). Ultimate causes and the evolution of altruism. *Behavioral Ecology and Sociobiology*, 65(3):503–512.

Marshall, J. A., Trimmer, P. C., Houston, A. I., and McNamara, J. (2013b). On evolutionary explanations of cognitive biases. *Trends in Ecology & Evolution*, 28(8):469–473.

Marshall, J. A., Trimmer, P. C., Houston, A. I., and McNamara, J. M. (2013c). On evolutionary explanations of cognitive biases. *Trends in Ecology & Evolution*, pages 469–473.

Masuda, N. (2012). Ingroup favoritism and intergroup cooperation under indirect reciprocity based on group reputation. *Journal of Theoretical Biology*, 311(0):8 – 18.

Masuda, N. and Nakamura, M. (2012). Coevolution of trustful buyers and cooperative sellers in the trust game. *PloS one*, 7(9):e44169.

Matsuo, T., Jusup, M., and Iwasa, Y. (2014). The conflict of social norms may cause the collapse of cooperation: Indirect reciprocity with opposing attitudes towards in-group favoritism. *Journal of Theoretical Biology*, 346(0):34 – 46.

Matthiesen, S. B., Bjørkelo, B., and Burke, R. J. (2011). Workplace Bully as the Dark Side of Whistleblowing. In Einarsen, S., Hoel, H., and Zapf, D., editors, *Bullying and Harassment in the Workplace: Developments in Theory, Research, and Practice*, pages 301–319.

McConnell, A. R., Dunn, E. W., Austin, S. N., and Rawn, C. D. (2011). Blind spots in the search for happiness: Implicit attitudes and nonverbal leakage predict affective forecasting errors. *Journal of Experimental Social Psychology*, 47(3):628 – 634.

McDermott, R., Tingley, D., Cowden, J., Frazzetto, G., and Johnson, D. D. P. (2009). Monoamine oxidase a gene (maoa) predicts behavioral aggression following provocation. *Proceedings of the National Academy of Sciences*, 106(7):2118–2123.

McDonald, S. and Ahern, K. (2000). The professional consequences of whistleblowing by nurses. *Journal of professional nursing : official journal of the American Association of Colleges of Nursing*, 16(6):313–21.

McKay, R. (2011). Isn't it ironic? a review of why everyone (else) is a hypocrite: Evolution and the modular mind. *Evolution and Human Behavior*, 32:444–446.

McKay, R. and Efferson, C. (2010). The subtleties of error management. *Evolution and Human Behavior*, 31(5):309–319.

McKay, R., Efferson, C., Whitehouse, H., and Fehr, E. (2011a). Wrath of god: religious primes and punishment. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1713):1858–1863.

McKay, R., Langdon, R., and Coltheart, M. (2005). "Sleights of mind": Delusions, defences, and self-deception. *Cognitive Neuropsychiatry*, 10(4):305–326.

McKay, R., Mijovic-Prelec, D., and Prelec, D. (2011b). Protesting too much: Self-deception and self-signaling. *Behavioral and Brain Sciences*, 34:34–35.

McKay, R. T. and Dennett, D. C. (2009). Our evolving beliefs about evolved misbelief. *Behavioral and Brain Sciences*, 32:541–561.

McKay, R. T., Dennett, D. C., et al. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, 32(6):493–510.

McNamara, J. M. and Houston, A. I. (2009). Integrating function and mechanism. *Trends in Ecology & Evolution*, 24(12):670–675.

McNamara, J. M., Stephens, P. A., Dall, S. R., and Houston, A. I. (2009a). Evolution of trust and trustworthiness: social awareness favours personality differences. *Proceedings of the Royal Society B: Biological Sciences*, 276(1657):605–613.

McNamara, J. M., Stephens, P. A., Dall, S. R. X., and Houston, A. I. (2009b). Evolution of trust and trustworthiness: social awareness favours personality differences. *Proceedings of the Royal Society B: Biological Sciences*, 276(1657):605–613.

McNamara, J. M. and Weissing, F. J. (2010). Evolutionary game theory. *Social behaviour: genes, ecology and evolution*, pages 88–106.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444.

Mele, A. R. (1997). Understanding and explaining real self-deception. *Behavioral and Brain Sciences*, 20:127–134.

Mele, A. R. (1999). Twisted self-deception. *Philosophical Psychology*, 12(2):117–137.

Mele, A. R. (2001). *Self-deception unmasked*. Cambridge Univ Press.

Mele, A. R. (2012). When are we self-deceived? *Humanamente*, 20:1–15.

Mellers, B. A. (2000). Choice and the relative pleasure of consequences. *Psychological bulletin*, 126(6):910.

Mesoudi, A. (2011). An experimental comparison of human social learning strategies: payoff-biased social learning is adaptive but underused. *Evolution and Human Behavior*, 32(5):334–342.

Meyvis, T., Ratner, R. K., and Levav, J. (2010). Why don't we learn to accurately forecast feelings? How misremembering our predictions blinds us to past forecasting errors. *Journal of Experimental Psychology: General*, 139(4):579–589.

Miceli, M. P. and Near, J. P. (2013). An International Comparison of the Incidence of Public Sector Whistle-Blowing and the Prediction of Retaliation: Australia, Norway, and the US. *Australian Journal of Public Administration*, 72(4):433–446.

Miceli, M. P., Near, J. P., and Dworkin, T. M. (2008). A Word to the Wise: How Managers and Policy-Makers can Encourage Employees to Report Wrongdoing. *Journal of Business Ethics*, 86(3):379–396.

Mijović-Prelec, D. and Prelec, D. (2009). Self-deception as self-signalling: a model and experimental evidence. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1538):227–240.

Millikan, R. G. (1989). In defense of proper functions. *Philosophy of Science*, 56(2):288–302.

Miloyan, B. and Suddendorf, T. (2015). Feelings of the future. *Trends in Cognitive Sciences*, 19(4):196 – 200.

Mitchell, D., Bryson, J. J., and Ingram, G. P. (2013). On the reliability of unreliable information: Gossip as cultural memory. *Unpublished*.

Mitchell, R. W. (1986). A framework for discussing deception. In Mitchell, R. W. and Thompson, N. S., editors, *Deception: Perspectives on human and nonhuman deceit*, pages 3–40. SUNY Press Albany, NY.

Mitzkewitz, M. and Nagel, R. (1993). Experimental results on ultimatum games with incomplete information. *International Journal of Game Theory*, 22(2):171–198.

Mokkonen, M. and Lindstedt, C. (2015). The evolutionary ecology of deception. *Biological Reviews*.

Moore, L. and McAuliffe, E. (2010). Is inadequate response to whistleblowing perpetuating a culture of silence in hospitals? *Clinical Governance: An International Journal*, 15(3):166–178.

Moore, M. T. and Fresco, D. M. (2012). Depressive realism: A meta-analytic review. *Clinical Psychology Review*, 32(6):496 – 509.

Morewedge, C. K. and Buechel, E. C. (2013). Motivated underpinnings of the impact bias in affective forecasts. *Emotion*, 13(6):1023–1029.

Morewedge, C. K., Gilbert, D. T., and Wilson, T. D. (2005). The least likely of times how remembering the past biases forecasts of the future. *Psychological Science*, 16(8):626–630.

Nakamaru, M. and Kawata, M. (2004). Evolution of rumours that discriminate lying defectors. *Evolutionary Ecology Research*, 6(2):261–283.

Nakamura, M. and Masuda, N. (2011). Indirect reciprocity under incomplete observation. *PLoS Comput Biol*, 7(7):e1002113.

Nakamura, M. and Masuda, N. (2012). Groupwise information sharing promotes ingroup favoritism in indirect reciprocity. *BMC Evolutionary Biology*, 12(1):213.

nas Garza, P. B., Espín, A. M., Exadaktylos, F., and Herrmann, B. (2014). Fair and unfair punishers coexist in the ultimatum game. *Scientific Reports*, 4(6025).

Near, J. P. and Miceli, M. P. (1985). Organizational dissidence: The case of whistle-blowing. *Journal of Business Ethics*, 4(1):1–16.

Nelkin, D. K. (2002). Self?deception, motivation, and the desire to believe. *Pacific Philosophical Quarterly*, 83(4):384–406.

Nelson, L. D. and Meyvis, T. (2008). Interrupted consumption: Disrupting adaptation to hedonic experiences. *Journal of Marketing Research*, 45(6):654–664.

New, J., Cosmides, L., and Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences*, 104(42):16598–16603.

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175.

Nisbett, R. E. and Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, 35(4):250–256.

Noë, R. and Hammerstein, P. (1994). Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology*, 35(1):1–11.

Novakova, J. and Flegr, J. (2013). How much is our fairness worth? the effect of raising stakes on offers by proposers and minimum acceptable offers in dictator and ultimatum games. *PLoS ONE*, 8(4):e60966.

Novemsky, N. and Ratner, R. K. (2003). The time course and impact of consumers' erroneous beliefs about hedonic contrast effects. *Journal of Consumer Research*, 29(4):507–516.

Nowak, M. A. and Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, 393:573–577.

Nowak, M. A. and Sigmund, K. (2004). Evolutionary dynamics of biological games. *science*, 303(5659):793–799.

Nowak, M. A. and Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437(7063):1291–1298.

Ohtsuki, H., Hauert, C., Lieberman, E., and Nowak, M. A. (2006). A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092):502–505.

Ohtsuki, H. and Iwasa, Y. (2004). How should we define goodness?reputation dynamics in indirect reciprocity. *Journal of theoretical biology*, 231(1):107–120.

Ohtsuki, H. and Iwasa, Y. (2006). The leading eight: social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, 239(4):435–444.

Orzack, S. H. and Forber, P. (2012). Adaptationism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition.

Oxoby, R. J. and McLeish, K. N. (2004). Sequential decision and strategy vector methods in ultimatum bargaining: Evidence on the strength of other-regarding behavior. *Economics Letters*, 84(3):399 – 405.

Pachur, T., Hertwig, R., and Wolkewitz, R. (2014). The affect gap in risky choice: Affect-rich outcomes attenuate attention to probability information. *Decision*, 1(1):64–78.

Panchanathan, K. (2011). Two wrongs don't make a right: The initial viability of different assessment rules in the evolution of indirect reciprocity. *Journal of Theoretical Biology*, 277(1):48 – 54.

Panchanathan, K. and Boyd, R. (2003). A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology*, 224(1):115–126.

Panhuis, T. M., Butlin, R., Zuk, M., and Tregenza, T. (2001). Sexual selection and speciation. *Trends in Ecology & Evolution*, 16(7):364–371.

Parker, G. A., Smith, J. M., et al. (1990). Optimality theory in evolutionary biology. *Nature*, 348(6296):27–33.

Perez-Truglia, R. (2012). On the causes and consequences of hedonic adaptation. *Journal of Economic Psychology*.

Peters, K., Luck, L., Hutchinson, M., Wilkes, L., Andrew, S., and Jackson, D. (2011). The emotional sequelae of whistleblowing: findings from a qualitative study. *Journal of clinical nursing*, 20(19-20):2907–14.

Peters, S. A., Laham, S. M., Pachter, N., and Winship, I. M. (2014). The future in clinical genetics: affective forecasting biases in patient and clinician decision making. *Clinical genetics*, 85(4):312–317.

Pfattheicher, S. and Keller, J. (2014). Towards a biopsychological understanding of costly punishment: The role of basal cortisol. *PLoS ONE*, 9(1):e85691.

Pfattheicher, S., Landhäußer, A., and Keller, J. (2014). Individual differences in antisocial punishment in public goods situations: The interplay of cortisol with testosterone and dominance. *Journal of Behavioral Decision Making*, 27(4):340–348.

Pfattheicher, S. and Schindler, S. (2015). Understanding the dark side of costly punishment: The impact of individual differences in everyday sadism and existential threat. *European Journal of Personality*, 29(4):498–505.

Pinker, S. (2005). A reply to jerry fodor on how the mind works. *Mind & Language*, 20(1):33–38.

Pinker, S. (2011). Representations and decision rules in the theory of self-deception. *Behavioral and Brain Sciences*, 34:35–37.

Pitt, M., Monks, T., Crowe, S., and Vasilakis, C. (2015). Systems modelling and simulation in health service design, delivery and decision making. *BMJ Quality & Safety*.

Porcher, J. E. (2012). Against the deflationary account of self-deception. *Humanamente*, 20:67–84.

Porcher, J. E. (2015). Can anosognosia vindicate traditionalism about self-deception? *Epistemology & Philosophy of Science*, 44(2):206–217.

Powers, S. T. and Lehmann, L. (2013). The co-evolution of social institutions, demography, and large-scale human cooperation. *Ecology Letters*, 16(11):1356–1364.

Powers, S. T., Penn, A. S., and Watson, R. A. (2011). The concurrent evolution of cooperation and the population structures that support it. *Evolution*, 65(6):1527–1543.

Powers, S. T., Taylor, D. J., and Bryson, J. J. (2012). Punishment can promote defection in group-structured populations. *Journal of Theoretical Biology*, 311:107–116.

Quattrone, G. A. and Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology*, 46(2):237–248.

Queller, D. C. and Strassmann, J. E. (2013). The veil of ignorance can favour biological cooperation. *Biology Letters*, 9(6).

Raihani, N. J. and Bshary, R. (2015). The reputation of punishers. *Trends in Ecology & Evolution*, 30(2):98–103.

Raihani, N. J., Mace, R., and Lamba, S. (2013). The effect of $1, $5 and $10 stakes in an online dictator game. *PLoS ONE*, 8(8):e73131.

Ramachandran, V. (1996). The evolutionary biology of self-deception, laughter, dreaming and depression: Some clues from anosognosia. *Medical Hypotheses*, 47(5):347 – 362.

Ramachandran, V. S. and Blakeslee, S. (1998). *Phantoms in the brain: Human nature and the architecture of the mind*. Fourth Estate, London.

Ramirez, J. and Marshall, J. (2015). Self-deception can evolve under appropriate costs. *Current Zoology*, 61(2):382–396.

Ramirez, M. K. (2007). Blowing the Whistle on Whistleblower Protection: A Tale of Reform versus Power. *University of Cincinnati Law Review*, 76(1).

Rand, D. G., Armao, J. J., Nakamaru, M., and Ohtsuki, H. (2010). Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology*, 265(4):624–632.

Rand, D. G., Greene, J. D., and Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416):427–430.

Rand, D. G., Tarnita, C. E., Ohtsuki, H., and Nowak, M. A. (2013). Evolution of fairness in the one-shot anonymous ultimatum game. *Proceedings of the National Academy of Sciences*, 110(7):2581–2586.

Rankin, D. J., dos Santos, M., and Wedekind, C. (2009). The evolutionary significance of costly punishment is still to be demonstrated. *Proceedings of the National Academy of Sciences*, 106(50):E135.

Rauwolf, P., Mitchell, D., and Bryson, J. J. (2015). Value homophily benefits cooperation but motivates employing incorrect social information. *Journal of Theoretical Biology*, 367(0):246–261.

Rayo, L. and Becker, G. S. (2007). Evolutionary efficiency and happiness. *Journal of Political Economy*, 115(2):302–337.

Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M., Fogarty, L., Ghirlanda, S., Lillicrap, T., and Laland, K. (2010). Why copy others? Insights from the social learning strategies tournament. *Science*, 328(5975):208.

Robson, A. and Samuelson, L. (2011). The evolution of decision and experienced utilities. *Theoretical Economics*, 6(3):311–339.

Ross, R. M., Greenhill, S. J., and Atkinson, Q. D. (2013). Population structure and cultural geography of a folktale in europe. *Proceedings of the Royal Society B: Biological Sciences*, 280(1756).

Rottenstreich, Y. and Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science*, 12(3):185–190.

Ruby, M. B., Dunn, E. W., Perrino, A., and Gillis, Randall, V.-S. (2011). The invisible benefits of exercise. *Health Psychology*, pages 67–74.

Santos, F. C., Santos, M. D., and Pacheco, J. M. (2008). Social diversity promotes the emergence of cooperation in public goods games. *Nature*, 454(7201):213–216.

Santos, L. R. and Rosati, A. G. (2015). The evolutionary roots of human decision making. *Annual Review of Psychology*, 66(1):321–347.

Sasaki, T., Okada, I., Uchida, S., and Chen, X. (2015). Commitment to cooperation and peer punishment: Its evolution. *Games*, 6(4):574.

Scheibe, S., Mata, R., and Carstensen, L. L. (2011). Age differences in affective forecasting and experienced emotion surrounding the 2008 us presidential election. *Cognition & emotion*, 25(6):1029–1044.

Schiestl, F. P. and Johnson, S. D. (2013). Pollinator-mediated evolution of floral signals. *Trends in Ecology & Evolution*, 28(5):307 – 315.

Schipper, L. D., Easton, J. A., and Shackelford, T. K. (2006). Morbid jealousy as a function of fitness-related life-cycle dimensions. *Behavioral and Brain Sciences*, 29:630–630.

Schkade, D. A. and Kahneman, D. (1998). Does living in california make people happy? a focusing illusion in judgments of life satisfaction. *Psychological Science*, 9(5):340–346.

Schmitt, P., Shupp, R., Swope, K., and Mayer, J. (2008). Pre-commitment and personality: Behavioral explanations in ultimatum games. *Journal of Economic Behavior & Organization*, 66(34):597 – 605.

Schonmann, R. H., Vicente, R., and Caticha, N. (2013). Altruism can proliferate through population viscosity despite high random gene flow. *PLoS ONE*, 8(8).

Schulz-Hardt, S., Frey, D., Lüthgens, C., and Moscovici, S. (2000). Biased information search in group decision making. *Journal of Personality and Social Psychology*, 78(4):655.

Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, 91(4):531–553.

Schwitzgebel, E. (2015). Belief. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Summer 2015 edition.

Sell, A., Cosmides, L., and Tooby, J. (2014). The human anger face evolved to enhance cues of strength. *Evolution and Human Behavior*, 35(5):425 – 429.

Sevdalis, N. and Harvey, N. (2009). Reducing the impact bias in judgments of post-decisional affect: Distraction or task interference? *Judgment and Decision Making*, 4(4):287–296.

Shah, A. K. and Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological bulletin*, 134(2):207–222.

Sharot, T. (2011). The optimism bias. *Current Biology*, 21(23):R941–R945.

Sharot, T., Riccardi, A. M., Raio, C. M., and Phelps, E. A. (2007). Neural mechanisms mediating optimism bias. *Nature*, 450(7166):102–105.

Sigmund, K. (2007). Punish or perish? retaliation and collaboration among humans. *Trends in Ecology & Evolution*, 22(11):593 – 600.

Sigmund, K. (2009). Sympathy and similarity: The evolutionary dynamics of cooperation. *Proceedings of the National Academy of Sciences*, 106(21):8405–8406.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63(2):129.

Simon, H. A. (1972). Theories of bounded rationality. In Radner, C. B. and Radner, R., editors, *Decision and Organization*, pages 161–176. North Holland, Amsterdam.

Simon, H. A. (1991). Bounded rationality and organizational learning. *Organization Science*, 2(1):pp. 125–134.

Skivenes, M. and Trygstad, S. (2015). Explaining whistle blowing processes in the Norwegian labour market: Between individual power resources and institutional arrangements. *Economic and Industrial Democracy*, pages 0143831X14559783–.

Skivenes, M. and Trygstad, S. C. (2010). When whistle-blowing works: The Norwegian case. *Human Relations*, 63(7):1071–1097.

Smith, D. L. (2004). Why we lie. the evolutionary roots of deception and the unconscious mind.

Smith, D. L. (2011). Aiming at self-deception: Deflationism, intentionalism, and biological purpose. *Behavioral and Brain Sciences*, 34:37–38.

Smith, D. L. (2014). Self-deception: A teleofunctional approach. *Philosophia*, 42(1):181–199.

Sommerfeld, R. D., Krambeck, H.-J., and Milinski, M. (2008). Multiple gossip statements and their effect on reputation and trustworthiness. *Proceedings of the Royal Society B: Biological Sciences*, 275(1650):2529–2536.

Stewart, A. J. and Plotkin, J. B. (2013). From extortion to generosity, evolution in the iterated prisoners dilemma. *Proceedings of the National Academy of Sciences*, 110(38):15348–15353.

Strack, F., Martin, L. L., and Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European journal of social psychology*, 18(5):429–442.

Suter, R. S., Pachur, T., and Hertwig, R. (2015). How affect shapes risky choice: Distorted probability weighting versus probability neglect. *Journal of Behavioral Decision Making*, pages n/a–n/a.

Suzuki, S. and Kimura, H. (2013). Indirect reciprocity is sensitive to costs of information transfer. *Scientific Reports*, 3.

Sylwester, K., Herrmann, B., and Bryson, J. J. (2013a). *Homo homini lupus?* Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics*, 6(3):167–188.

Sylwester, K., Mitchell, J., and Bryson, J. J. (2013b). Punishment as aggression: Uses and consequences of costly punishment across populations. to be resubmitted.

Sylwester, K. and Roberts, G. (2010). Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters*.

Sylwester, K. and Roberts, G. (2013). Reputation-based partner choice is an effective alternative to indirect reciprocity in solving social dilemmas. *Evolution and Human Behavior*, 34(3):201–206.

Takahashi, N. and Mashima, R. (2006). The importance of subjectivity in perceptual errors on the emergence of indirect reciprocity. *Journal of Theoretical Biology*, 243(3):418 – 436.

Tanabe, S., Suzuki, H., and Masuda, N. (2013). Indirect reciprocity with trinary reputations. *Journal of Theoretical Biology*, 317:338–347.

Tarnita, C. E. (2015). Fairness and trust in structured populations. *Games*, 6(3):214.

Tavakoli, A. A., Keenan, J. P., and Cranjak-Karanovic, B. (2003). Culture and whistleblowing an empirical study of croatian and united states managers utilizing hofstede's cultural dimensions. *Journal of Business Ethics*, 43:49–64.

Taylor, P. D., Day, T., and Wild, G. (2007). Evolution of cooperation in a finite homogeneous graph. *Nature*, 447(7143):469–472.

Taylor, S. and Brown, J. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, 103(2):193–210.

Taylor, S. E., Lerner, J. S., Sherman, D. K., Sage, R. M., and McDowell, N. K. (1983). Portrait of the self-enhancer: Well adjusted and well liked or maladjusted and friendless? *American Psychologist*, 84(1):165–176.

Tella, R. D., New, J. H.-D., and MacCulloch, R. (2010). Happiness adaptation to income and to status in an individual panel. *Journal of Economic Behavior & Organization*, 76(3):834 – 852.

Tenney, E. R., Spellman, B. A., and MacCoun, R. J. (2008). The benefits of knowing what you know (and what you don?t): How calibration affects credibility. *Journal of Experimental Social Psychology*, 44(5):1368 – 1375.

Todd, P. M. and Gigerenzer, G. (2000). Precis of simple heuristics that make us smart. *Behavioral and Brain Sciences*, 23:727–741.

Traag, V. A., Van Dooren, P., and Nesterov, Y. (2011). Indirect reciprocity through gossiping can lead to cooperative clusters. In *Artificial Life (ALIFE), 2011 IEEE Symposium on*, pages 154–161. IEEE.

Traulsen, A., Pacheco, J. M., and Nowak, M. A. (2007). Pairwise comparison and selection temperature in evolutionary game dynamics. *Journal of Theoretical Biology*, 246(3):522 – 529.

Trimmer, P. C., Higginson, A. D., Fawcett, T. W., McNamara, J. M., and Houston, A. I. (2015). Adaptive learning can result in a failure to profit from good conditions: implications for understanding depression. *Evolution, Medicine, and Public Health*, 2015(1):123–135.

Trimmer, P. C., Houston, A. I., Marshall, J. A., Bogacz, R., Paul, E. S., Mendl, M. T., and McNamara, J. M. (2008). Mammalian choices: combining fast-but-inaccurate and slow-but-accurate decision-making systems. *Proceedings of the Royal Society of London B: Biological Sciences*, 275(1649):2353–2361.

Trimmer, P. C., Houston, A. I., Marshall, J. A., Mendl, M. T., Paul, E. S., and McNamara, J. M. (2011). Decision-making under uncertainty: biases and bayesians. *Animal cognition*, 14(4):465–476.

Trimmer, P. C., Marshall, J. A., Fromhage, L., McNamara, J. M., and Houston, A. I. (2013). Understanding the placebo effect from an evolutionary perspective. *Evolution and Human Behavior*, 34(1):8 – 15.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46(1):35–57.

Trivers, R. (1985). *Social evolution*. Benjamin/Cummings Publishing Company Menlo Park.

Trivers, R. (1991). Deceit and self-deception: The relationship between communication and consciousness. *Man and Beast Revisited*, pages 175–191.

Trivers, R. (2000). The elements of a scientific theory of self-deception. *Annals of the New York Academy of Sciences*, 907(1):114–131.

Trivers, R. (2011). *The folly of fools: The logic of deceit and self-deception in human life*. Basic Books.

Trongmateerut, P. and Sweeney, J. T. (2012). The influence of subjective norms on whistle-blowing: A cross-cultural investigation. *Journal of Business Ethics*, 112(3):437–451.

Turner, M. E. and Pratkanis, A. R. (1998). Twenty-five years of groupthink theory and research: Lessons from the evaluation of a theory. *Organizational Behavior and Human Decision Processes*, 73(2):105–115.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211:453–58.

Ubel, P. A., Loewenstein, G., and Jepson, C. (2005). Disability and sunshine: Can hedonic predictions be improved by drawing attention to focusing illusions or emotional adaptation? *Journal of Experimental Psychology: Applied*, 11(2):111–123.

Uchida, S. and Sasaki, T. (2013). Effect of assessment error and private information on stern-judging in indirect reciprocity. *Chaos, Solitons & Fractals*, 56:175–180.

Uchida, S. and Sigmund, K. (2010). The competition of assessment rules for indirect reciprocity. *Journal of Theoretical Biology*, 263(1):13–19.

Vaillant, G. E. (1992). *Ego Mechanisms of Defense: A Guide for Clinicans and Researchers*. American Psychiatric Pub.

Van Valen, L. (1973). A new evolutionary law. *Evolutionary theory*, 1:1–30.

Varki, A. and Brower, D. (2013). *Denial: Self-Deception, False Beliefs, and the Origins of the Human Mind*. Hachette Digital, Inc.

von Hippel, W. (2015). Self-deception. In Zeigler-Hill, V., Welling, L. L. M., and Shackelford, T. K., editors, *Evolutionary Perspectives on Social Psychology*, Evolutionary Psychology, pages 149–158. Springer International Publishing.

von Hippel, W. and Trivers, R. (2011a). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, 34(1):1.

von Hippel, W. and Trivers, R. (2011b). Reflections on self-deception. *Behavioral and Brain Sciences*, 34:41–56.

Vrij, A. (2011). Self-deception, lying, and the ability to deceive. *Behavioral and Brain Sciences*, 34:40–41.

Wallace, B., Cesarini, D., Lichtenstein, P., and Johannesson, M. (2007). Heritability of ultimatum game responder behavior. *Proceedings of the National Academy of Sciences*, 104(40):15631–15634.

Watson, C. L. and OConnor, T. (2015). Legislating for advocacy: The case of whistleblowing . *Nursing Ethics*.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, 39(5):806–820.

Weinstein, N. D. (1982). Community noise problems: Evidence against adaptation. *Journal of Environmental Psychology*, 2(2):87 – 97.

Wenze, S. J., Gunthert, K. C., Ahrens, A. H., and Taylor Bos, T. C. (2013). Biases in short-term mood prediction in individuals with depression and anxiety symptoms. *Individual Differences Research*, 11(2):91 – 101.

Whiten, A., McGuigan, N., Marshall-Pescini, S., and Hopper, L. M. (2009). Emulation, imitation, overimitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1528):2417–2428.

Wilke, A. and Barrett, H. C. (2009). The hot hand phenomenon as a cognitive adaptation to clumped resources. *Evolution and Human Behavior*, 30(3):161 – 169.

Williams, L. E. and Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science*, 322(5901):606–607.

Williamson, O. E. (1993). Calculativeness, trust, and economic organization. *The Journal of Law & Economics*, 36(1):453–486.

Willner-Reid, J., Smith, N., Jones, H. B., and MacLeod, A. K. (2012). Affective forecasting in problem gamblers. *International Gambling Studies*, 12(3):295–307.

Wilson, T. D. and Gilbert, D. T. (2003). Affective forecasting. *Advances in experimental social psychology*, 35:345–411.

Wilson, T. D. and Gilbert, D. T. (2013). The impact bias is alive and well. *Journal of Personality and Social Psychology*, 105(5):740–748.

Wilson, T. D., Gilbert, D. T., and Centerbar, D. B. (2003a). Making sense: The causes of emotional evanescence. *The psychology of economic decisions*, 1:209–233.

Wilson, T. D., Meyers, J., and Gilbert, D. T. (2003b). How happy was I, anyway? A retrospective impact bias. *Social Cognition*, 21(6):421–446.

Wilson, T. D., Wheatley, T., Meyers, J. M., Gilbert, D. T., Axsom, D., et al. (2000). Focalism: A source of durability bias in affective forecasting. *Journal of personality and social psychology*, 78(5):821–836.

Wolf, Max hand McNamara, J. M. (2013). Adaptive between-individual differences in social competence. *Trends in ecology & evolution*, 28(5):253–254.

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., Miura, A., Inukai, K., Takagishi, H., and Simunovic, D. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences*, 109(50):20364–20368.

Zhao, K. and Smillie, L. D. (2014). The role of interpersonal traits in social decision making: Exploring sources of behavioral heterogeneity in economic games. *Personality and Social Psychology Review*.

Zitzler, E. and Thiele, L. (1998). Multiobjective optimization using evolutionary algorithms?a comparative case study. In *Parallel problem solving from nature?PPSN V*, pages 292–301. Springer.